МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ОДЕСЬКИЙ ДЕРЖАВНИЙ ЕКОЛОГІЧНИЙ УНІВЕРСИТЕТ

М.І. Бургаз

СТАТИСТИЧНІ МЕТОДИ В БІОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

Конспект лекцій

Одеса
Одеський державний екологічний університет
2024

УДК 519.2
Б 90

Бургаз М.І.
Б 90 Статистичні методи в біологічних дослідженнях : конспект лекцій. Одеса: ОДЕКУ, 2024. 114с.
ISBN 978-966-186-321-6

В конспекті лекцій висвітлені питання, щодо основних математико-статистичних методів, що застосовуються в біологічних дослідженнях. Розкриті основні питання закономірностей розподілу випадкових величин, методика побудови варіаційних рядів, характеристики законів розподілу випадкових величин та оцінки вибіркових показників. Викладаються основні аспекти кореляційного, регресійного, дисперсійного аналізу.

Конспект лекцій для студентів IV курсу денної форми навчання за спеціальністю "Водні біоресурси та аквакультура".

Конспект лекцій для студентів рівня вищої освіти «Бакалавр» IV-V років навчання денної та заочної форм навчання за спеціальністю 207 "Водні біоресурси та аквакультура".

УДК 519.2

Рекомендовано методичною радою Одеського державного екологічного університету Міністерства освіти і науки України як конспект лекцій (протокол № 5 від 25. 04. 2024 р.)

ISBN 978-966-186-321-6

ЗМІСТ

ВСТУП

Конспект лекцій підготовлено відповідно до силлабусу дисципліни «Статистичні методи в біологічних дослідженнях», що входить до складу дисциплін з підготовки фахівців рівня вищої освіти «Бакалавр» спеціальності 207 «Водні біоресурси і аквакультура».

Біологічна статистика як самостійна наукова дисципліна виникла в XIX столітті. Однак її витоки сягають до більш раннього періоду історії природознавства, до того часу, коли вимірювання стали розглядати як один з методів пізнання природи.

Значення біологічної статистики у дослідницькій роботі і в професійній підготовці біологів і фахівців суміжних галузей знання стало очевидним уже тоді, коли були відкриті статистичні закони, що діють у сфері масових явищ. Але біологи не відразу оцінили всю важливість цих відкриттів.

Сучасна біологічна статистика – велика область знань, в ній нерозривно пов'язані питання планування біологічних експерименту та методика статистичного аналізу їх результатів. Зневажливе ставлення до біологічної статистики, ігнорування її вимог і правил призводить не лише до зайвих витрат часу і праці на виконання дослідницької роботи, але нерідко і до недостатньо обґрунтованих і помилкових висновків.

Системний підхід до аналізу складних явищ – одна з провідних ідей сучасного природознавства, в тому числі і біології. Ця ідея і пов'язані з нею способи моделювання складних явищ зачіпають і самий спосіб мислення. Біологічна статистика має безпосереднє відношення до всіх цих фактів. Спираючись на закони, що діють у статистичних сукупностях, біологічна статистика озброює дослідника не тільки потрібними знаннями в області статистичного аналізу масових явищ, але і виховує у біолога статистичне мислення, розширюючи тим самим його науковий кругозір. Біологічна статистика розкриває перед нами діалектику зв'язку між частиною і цілим, між поодинокими фактами та їх сукупністю, між причиною і наслідком, випадковим і необхідним у явищах живої природи. Вона показує, що в уявному хаосі випадковостей проявляються закономірності, доступні опису точними математичними методами.

Біологічна статистика служить об'єктивною основою порівняльного методу, без якого взагалі неможливе пізнання реальної дійсності. Воістину важко переоцінити ту роль, яку покликана відігравати біологічна статистика у розвитку біології та суміжних наук, і можна не сумніватися, що в ході подальших успіхів природознавства роль біологічної статистики у дослідницькій роботі і в професійній підготовці біологів різних профілів буде зростати.

Біологічна статистика вносить вагомий внесок до скарбниці наших знань про природу і про способи її пізнання. При цьому вона не зазіхає на історично сформовані і до цих пір виправдовувані себе описові методи дослідження, не заперечує і не підміняє їх, а лише озброює дослідника ідеями і методами, які застосовуються при вивченні варіюючих об'єктів.

Біологічна статистика – наука формальна і застосовувати її треба вміло, з урахуванням специфіки досліджуваних явищ.

Тільки добре засвоївши теоретичні основи біологічної статистики, можна розраховувати на успішне застосування її методів у дослідницькій роботі. При цьому не слід допускати надмірностей, жонглювання біометричними показниками, виставляючи їх напоказ там, де цього не потребує суть справи. Не слід також вдаватися в іншу крайність: обмежуватися примітивними способами аналізу біометричних даних. Істина полягає не в крайнощах, а в розумному підході до справи. Невміле застосування біологічної статистики може принести не користь, а шкоду.

При підготовці цього конспекту лекцій були використані літературні джерела довідкового характеру, посібники та підручники вітчизняних та іноземних авторів.

1 ВІДМІТНІ РИСИ БІОЛОГІЧНОЇ СТАТИСТИКИ ТА ЇЇ МІСЦЕ В СИСТЕМІ БІОЛОГІЧНИХ НАУК

Біологічна статистика склалася в примежових областях між біологією та математикою. Її розвиток пов'язаний з перетворенням біології з науки описової в науку точну, засновану на вимірюваннях, на застосуванні кількісних оцінок при вирішенні біологічних завдань. З формальної точки зору біологічна статистика представляє «сукупність математичних методів, що застосовуються у біології». Ці методи вона запозичує головним чином з області математичної статистики і теорії ймовірностей.

Математична статистика і теорія ймовірностей – розділи математики; це науки суто теоретичні, розглядають масові явища в абстрактній формі, незалежно від їх конкретного змісту. Теорія ймовірностей вивчає закони поведінки випадкових подій і випадкових змінних величин, а математична статистика займається розробкою теорії вибіркового методу. Біологічна статистика ж – наука прикладна, що має справу з конкретними фактами, які вона аналізує за допомогою методів математичної статистики і теорії ймовірностей, що становлять у сукупності те, що називають **статистичним аналізом**. Біологічна статистика ставить не математичні, а виключно біологічні цілі, використовуючи математико-статистичні методи під власним кутом зору, пристосовуючи їх до задач і специфіки біологічних досліджень. *Біологічна статистика* – це розділ біології, а не математики, вона має свій предмет і займає певне положення в системі біологічних наук.

Зв'язки сучасної біології з математикою багатобічні, вони все більше розширюються і поглиблюються. На стику біології з математикою виникли різні напрямки математичної біології. Кожен напрямок має свої завдання і стосовно до них використовує відповідні математичні методи. Біологічна статистика хоч і тісно пов'язана з математичної біологією, але ототожнюватися з нею не повинна, Математична біологія підходить до вирішення біологічних проблем дедуктивно, висуваючи на перший план математичні моделі з наступною перевіркою їх досвідом. Біологічна статистика ж спирається на індуктивний метод, відправляючись від конкретних фактів, які вона аналізує методами математичної статистики і теорії ймовірностей.

Сучасна біологічна статистика – це розділ біології, змістом якого є планування спостережень і статистичний аналіз їх результатів. Причому під спостереженням в широкому сенсі мається на увазі процес планомірного добування та накопичення фактичних даних незалежно від того, як він здійснюється – в експерименті або безпосереднім описом

досліджуваних явищ. Планування спостережень і обробка їх результатів – нерозривно пов'язані завдання статистичного аналізу.

1.1 Предмет і основні поняття біологічної статистики

Біологічна статистика – наука про статистичний аналіз масових явищ в біології, тобто таких явищ, в масі яких виявляються закономірності, які не виявляються на одиничних випадках спостережень. Наприклад, рибак ловив рибу та записав кількість виловленої риби в журнал – це окремий випадок, одиничне явище. Якщо ж риболовлю проводили одночасно декілька рибаків – це явище масове незалежно від того, яким був об'єкт спостереження – одиничним або груповим.

Предметом біологічної статистики служить будь-який біологічний об'єкт, якщо спостереження, які над ним проводяться отримують кількісну оцінку. Зазвичай спостереження проводяться не на одиничних, а на групових об'єктах, наприклад на особинах одного і того ж виду, статі і віку, які розглядаються в якості складових елементів або членів групового об'єкта і називаються **одиницями спостереження**. Сукупність таких відносно однорідних, але індивідуально помітних одиниць, що об'єднуються у відношенні деяких загальних умов для спільного (групового) вивчення, називається **статистичною сукупністю**.

Поняття статистичної сукупності – одне з фундаментальних понять біологічної статистики, воно базується на принципі якісної однорідності її складу. Не можна вивчати закономірність модифікацій (неспадкових змін фенотипу) на генетично неоднорідному матеріалі і т. д.

Статистична сукупність може складатися не тільки з аморфної маси однорідних одиниць, але й з різних за складом, але внутрішньо однорідних груп (особин, клітин і т, п.), що об'єднуються щодо прийнятих в досвіді умов для спільної обробки. У таких випадках сукупність вихідних даних називається **статистичним комплексом**. Питання про структуру статистичної сукупності вирішується дослідником в залежності від об'єкта і мети дослідження. Якої б форми і змісту не набирала статистична сукупність, вона завжди представляє певну систему, що не зводиться до арифметичної суми складових її одиниць та компонентів. У статистичних сукупностях існує внутрішній зв'язок між частиною і цілим, одиничним і загальним, який і знаходить своє вираження у статистичних законах, що діють у сфері масових явищ. На ці закони спирається біологічна статистика.

1.2 Ознаки, їх властивості та класифікація

Спостереження над біологічними об'єктами проводять за тими або іншими ознаками, тобто за такими характерними особливостями в будові і функціях живого, за якими можна відрізнити одну одиницю спостереження від іншої, порівнювати їх між собою.

Всі біологічні ознаки варіюють, тобто змінюються від випадку до випадку в певних межах. **Варіювання** – характерна властивість всього живого. Не треба мати виняткову спостережливість для того, щоб в масі однорідних особин бачити більш-менш помітні індивідуальні відмінності у величині, забарвленні, поведінці та інших ознаках і властивостях індивідів. Вимірюючи вміст жиру в рибі, підраховуючи кількість ікринок, зважуючи мальків одного і того ж посліду, – у всіх таких і подібних випадках величина кожної ознаки буде коливатися в деяких межах від однієї одиниці спостереження до іншої. Ці коливання величини однієї і тієї ж ознаки, що спостерігаються в загальній масі його числових значень, називаються **варіаціями**, а окремі числові значення варіюючого ознаки прийнято називати **варіантами** (від лат. *Variants, variantis* – помітний, що змінюється).

Всі біологічні ознаки варіюють, але не всі піддаються безпосередньому виміру. Звідси випливає їх розподіл на ***якісні***, або ***атрибутивні, і кількісні***. Якісні ознаки не піддаються безпосередньому виміру і обліковуються за наявності їх у членів даної сукупності. Наприклад, в масі риб неважко відрізнити і врахувати кількість особин різної статі або різної масті. Кількісні ознаки, такі, наприклад, як розмір і маса самців і самок риб, у кількості наявних у них ікринок, і т. п., можна безпосередньо виміряти або порахувати.

Розподіл ознак на якісні та кількісні умовно (хоча і необхідно з точки зору біологічної статистики): у кожному ролі можна виявити безліч градацій кількісних, наприклад у забарвленні (за кількістю міститься в них пігменту), так само як і сукупність числових значенні кількісних ознак можна підрозділити на якісно відособлені групи, наприклад, на хороших, посередніх і поганих і т.д.

Біологічні ознаки можна класифікувати по-різному залежно від того, що приймається за основу класифікації. Якщо основу класифікації становить той чи інший спосіб групування біометричних даних, ознаки ділять на альтернативні, порядкові, рангові і т.д. Загальною ж основою для класифікації ознак (при кількісному підході до їх опису) служить в одних випадках міра, в інших – рахунок. Це відноситься не тільки до кількісних, але і до якісних ознаками. На цій підставі біологічні ознаки поділяються на

мірні, або ***метричні,*** і ***лічильні,*** або ***меристичні***. Мірні ознаки варіюють безперервно: їх величина може в певних межах (від – до) приймати будь-які числові значення, тоді як лічильні ознаки варіюють переривчасто або дискретно – їх значення завжди виражаються тільки цілими числами.

Мовою математики величина будь-якої варіюючої ознаки називається **змінною випадкової величини**. На відміну від постійних величин, що позначаються початковими буквами латинського алфавіту, змінні величини прийнято позначати останніми в латинському алфавіті прописними буквами $X, Y, Z, \ldots$, а нечислові значення, тобто варіанти, – відповідними малими літерами того ж алфавіту $x_1, x_2, x_3, \ldots, x_n$ або $y_1, y_2, y_3, \ldots, y_n$ й т.д.

Загальне позначення будь-якої варіанти відзначається символом x_i, y_i або z_i, і т. д., де значок i вказує на порядковий номер варіанти.

Оскільки ознаки, за якими проводяться спостереження, виражаються тими чи іншими одиницями виміру або одиницями розрахунку, поняття статистичної сукупності поширюється не тільки на групи однорідних натуральних одиниць, але і на масу числових значень ознаки або ознак, за якими ця сукупність утворена.

1.2.1 Джерела варіювання ознак

Ознаки варіюють під впливом різних, в тому числі і численних випадкових причин. На результатах ж спостережень позначаються ще й помилки, допущені при вимірах. Досвід показав, що вимірювання, як би точно вони не вироблялися, завжди супроводжуються більш або менш помітними похибками. Ці похибки, або помилки, виникають через несправність або недостатньої точності вимірювальних приладів та інструментів (технічні помилки), від особистих якостей дослідника, його навичок та майстерності в роботі (особисті помилки) і від низки інших, не піддаються регулюванню і непереборних причин (випадкові помилки).

Технічні та особисті помилки, що об'єднуються в категорію систематичних, тобто невипадкових, помилок, можна значною мірою подолати, удосконалюючи технічні засоби, умови спостережень і особистий досвід. Ці заходи дозволяють звести розміри таких помилок до мінімуму, яким можна знехтувати. Випадкові ж помилки залишаються і разом з природним варіюванням позначаються на результатах спостережень. Однак у порівнянні з природним варіюванням ознак випадкові помилки вимірів, як правило, невеликі. Тому варіювання

результатів спостережень розглядається звичайно як варіювання досліджуваних ознак.

1.3 Точність вимірювань та правила округлювання дробових чисел

Зазвичай біологічні ознаки вимірюють з точністю до десятих, сотих йди тисячних часток одиниці, рідше виробляються більш точні вимірювання. Практично кожна ознака має свою міру.

Спираючись на кількісні показники, дослідник, як правило, має справу з наближеними величинами, йому доводиться постійно вдаватися до округлення дробових чисел. При цьому слід дотримуватися наступних правил:

1. якщо цифра, що стоїть за тою, що зберігається, менше 5, то вона відкидається, а якщо більше 5, то цифра, що зберігається збільшується на одиницю. Наприклад, округлюючи числа 45,346 і 8,644 до сотого знаку, отримуємо 45,35 і 8,644;
2. коли необхідно відкинути цифру 5, за якою немає інших цифр, то останню цифру залишають без зміни, якщо вона парна, і збільшують на одиницю, якщо вона непарна. Наприклад, числа 3,585 та 3,575, що округлюються до сотого знаку, дають одну і ту ж величину – 3,58. Припустимо і інше правило: округляти такі числа тільки в бік збільшення. Слідуючи цьому правилу, числа 3,585 і 3,575 округлюють до сотого знаку: 3,59 і 3,58.

Питання для самоперевірки:

1. Що таке сучасна біологічна статистика?
2. Що є предметом біологічної статистики?
3. Що називається статистичною сукупністю?
4. Що називається одиницями спостережень біологічної статистики?
5. Що таке статистичний комплекс?
6. Що таке варіації?
7. Що таке варіанти?
8. Назвіть основні біологічні ознаки статистики?
9. Як відбувається класифікація біологічних ознак?
10. Назвіть основні правила округлення дробових чисел.

2 ГРУПУВАННЯ ПЕРВИННИХ ДАНИХ

2.1 Генеральна сукупність і вибірка

Спостереження, що проводяться над біологічними об'єктами, можуть охоплювати всіх членів досліджуваної сукупності без винятку і можуть обмежуватися обстеженням лише деякої частини членів даної сукупності. У першому випадку спостереження буде називатися **суцільним** або **повним**, а в другому – **частковим** чи **вибірковим**. Суцільне спостереження дозволяє отримувати вичерпну інформацію про груповий об'єкті, у чому й полягає перевага цього способу перед способом вибіркового спостереження. Однак до суцільного спостереження вдаються не завжди. *По-перше*, тому що ця робота пов'язана з великими витратами часу та праці, а *по-друге*, в силу практичної неможливості або недоцільності проведення такої роботи. Неможливо, наприклад, врахувати всіх мешканців зоо- чи фітопланктону навіть невеликої водойми, так як їх чисельність практично неозора. Тому в переважній більшості випадків замість суцільного спостереження вивченню піддають якусь частину обстежуваної сукупності, за якою і судять про її стан в цілому.

Сукупність, з якої відбирається деяка частина її членів для спільного вивчення, називається **генеральною**, а відібрана тим чи іншим способом частина генеральної сукупності отримала назву **вибіркової сукупності** або **вибірки**. Обсяг генеральної сукупності, що позначається літерою N, теоретично нічим не обмежений ($N \to \infty$), тобто генеральна сукупність мислиться як нескінченно велика безліч відносно однорідних одиниць чи членів, що складають її зміст. Практично ж обсяг генеральної сукупності завжди обмежений і може бути різним, що залежить як від об'єкта спостереження, так і від завдання, поставленого перед дослідником.

Обсяг вибірки, що позначається літерою n, може бути і порівняно великим і малим, але він не може містити менше двох одиниць. Вибірковий метод є основним при вивченні статистичних сукупностей. Його перевага перед суцільним урахуванням всіх членів генеральної сукупності полягає в тому, що він скорочує час і витрати праці (за рахунок зменшення числа спостережень), а головне – дозволяє отримувати інформацію про такі сукупності, суцільне обстеження яких практично неможливо або недоцільно.

2.2 Репрезентативність вибірки

Основне завдання, що вирішується за допомогою вибіркового методу, зводиться до отримання такої інформації, яка дозволяє більш-менш точно судити про стан генеральної сукупності. Досвід показав, що вибірка досить добре відображає структуру генеральної сукупності. Однак, як правило, повного збігу вибіркових характеристик з характеристиками генеральної сукупності не буває. Щоб вибірка якнайповніше відображала структуру генеральної сукупності, вона повинна бути досить представницької, або **репрезентативною** (від лат. *Represento* – уявляю). Репрезентативність вибірки досягається способом **рендомізації** (від англ. *Random* – випадок), або випадковим відбором варіант з генеральної сукупності, що забезпечує рівну можливість для всіх членів генеральної сукупності потрапити до складу вибірки. Існує два основних способи відбору варіант з генеральної сукупності: **повторний** і **безповторний**. *Повторний відбір* проводиться за схемою «повернення» врахованих одиниць в генеральну сукупність, так що одна і та ж одиниця може потрапити у вибірку повторно. Цей відбір не впливає на склад генеральної сукупності і можливість кожній одиниці потрапити до вибірки не змінюється. При *безповторному* відборі враховані одиниці в генеральну сукупність не повертаються, кожна відібрана одиниця реєструється один раз. Можливість одиниць генеральної сукупності потрапити до вибірки змінюється при безповторному відборі, оскільки кожен попередній відбір впливає на результати подальшого і на склад генеральної сукупності, який теж змінюється. У практиці застосовується зазвичай безповторний випадковий відбір. Випадковий повторний відбір є теоретичною моделлю, що дозволяє дослідити процеси, які відбуваються у статистичних сукупностях, що має велике пізнавальне значення.

Ідеальний випадковий відбір, як повторний, так і безповторний, проводиться за способом жеребкування або лотереї, а також за допомогою таблиці випадкових чисел, що дозволяє повністю виключити суб'єктивні впливи на склад вибірки. Сутність цього способу полягає в наступному. На чисельно обмеженій, але досить великій штучній моделі генеральної сукупності методом повторного випадкового відбору утворюється ряд чисел, які заносяться в таблицю таким чином, щоб вони мали однакову кількість цифр. Цим полегшується зручність використання такої таблиці в практичних цілях.

Принцип рендомізації не виключає плановості відбору одиниць з генеральної сукупності і може здійснюватися по-різному залежно від

завдання і організації експерименту. Розрізняють такі види планового відбору:

1) типовий, або груповий,
2) серійний, або гніздовий,
3) механічний.

При ***типовому відборі*** генеральна сукупність попередньо ділиться на типові групи (ділянки, райони). Потім у кожній групі випадковим способом відбирається однакове або пропорційне число одиниць, що об'єднуються потім в одну вибіркову сукупність, яка і піддається статистичному аналізу. У випадках ***серійного відбору*** генеральна сукупність, як і при типовому відборі, попередньо ділиться на групи, звані гніздами або серіями. Потім, на розсуд дослідника, із загальної кількості серій відбирається необхідне їх число для спільної обробки. При цьому серії можуть бути одно- і різночисленними. Таким чином при серійному відборі на відміну від відбору типового з генеральної сукупності відбираються не окремі одиниці, а цілі серії або гнізда щодо однорідних одиниць. ***Механічний відбір*** здійснюється за такою схемою. Генеральна сукупність розбивається на декілька рівних частин або груп. Потім з кожної групи випадковим способом відбирають по одній одиниці. Таким чином, при механічному відборі число відібраних одиниць дорівнює числу груп, на які розбита генеральна сукупність.

В інших випадках механічний відбір проводиться з кожної десятки, сотні і т.д. зустрічаємих одиниць генеральної сукупності. Наприклад, при проведенні ботанічних або зоологічних екскурсій у вибірку потрапляє кожен 10-й, 20-й примірник і т.д. витрачених рослин або тварин даного виду.

Щоб вибірка була найбільш репрезентативною, необхідно поряд з правильно організованим відбором варіант звертати увагу на розмах варіювання ознаки і погоджувати з ним обсяг вибірки. Чим ширше розмах варіювання ознаки, тим більшим має бути і обсяг вибірки. Нечисленний склад вибірки при сильному варіюванні ознаки знижує її репрезентативність.

2.3 Групування первинних даних

Спостереження над біологічними об'єктами проводяться одночасно за кількома ознаками, що дозволяє зібрати найбільш повні відомості. Результати спостережень фіксуються в щоденниках, протоколах, анкетах, бланках та в інших формулярах первинного обліку. В умовах лабораторного експерименту результати випробувань фіксують в

протоколах, журналах, бланках і т.п. Форм і способів обліку результатів спостережень багато. Первинні документи обліку містять фактичний матеріал, що потребує обробки. Обробка починається з упорядкування зібраних даних (дотримуючись правила однорідності складу вибірки), систематизації виражених числами фактів, з тим щоб витягти укладену в них інформацію. Процес систематизації, або впорядкування, первинних біометричних даних з метою отримання укладеної в них інформації, виявлення закономірності, якій підлягає досліджуване явище чи процес, називається **групуванням**.

Групування – це не просто технічний прийом, а глибоко осмислена дія, спрямована на отримання правдивої і повноцінної інформації про досліджуваному об'єкті. Обраний спосіб групування повинен відповідати вимогу поставленого завдання і добре узгоджуватися з вмістом досліджуваного явища.

Групування вихідних даних може бути різною в залежності від того, з якою метою і за якими ознаками вона проводиться. Найбільш прийнятною формою групування є статистичні таблиці. Зазвичай в таблицях наводяться і загальні підсумки – у вигляді сум або усереднених показників, а також у відсотках від чисельності варіант у групах і у всій групування в цілому.

Групування за однією ознакою називається **простою**, а за кількома ознаками – **складною**. Звідси і таблиці можуть бути простими і складними.

Не менш складною виявляється групування вибіркових даних при з'ясуванні зв'язку між варіюючими ознаками. У таких випадках числові значення ознак з урахуванням їх повторюваності в димерній сукупності групуються в ***кореляційну таблицю***.

Особливий інтерес для біолога представляє угрупування вихідних даних у статистичні ряди – ряди числових значень ознаки, розташованих у певному порядку. У залежності від того, в якому плані (статики або динаміки) і за якими ознаками (якісним або кількісним) розглядаються явища або процеси, статистичні ряди поділяються на **атрибутивні**, **варіаційні**, **динаміки** або **тимчасові**.

Найбільше значення в курсі біологічної статистики мають варіаційні ряди. **Варіаційним рядом** називається ряд чисел, що показує закономірність розподілу одиниць досліджуваної сукупності за ранжованим значенням варіюючої ознаки (від франц. *Ranger* – вибудовувати в ряд по ранжиру, тобто по росту). Цей подвійний ряд чисел, що показує, яким чином числові значення ознаки (x_i) пов'язані з їх повторюваністю (m_i) в даній сукупності, називається **варіаційним** або **рядом розподілу**. Числа, що показують, скільки разів окремі варіанти зустрічаються в даній сукупності, називаються **частотами** або **вагами варіант** і позначаються m. Загальна сума частот завжди дорівнює обсягу

даної сукупності, тобто $\sum m = n$, де $\sum$ – підсумовування частот варіаційного ряду, n – обсяг вибіркової сукупності.

Частоти виражаються не тільки абсолютними, але і відносними числами – в частках одиниці або у відсотках від загальної чисельності варіант, що складають дану сукупність. У таких випадках ваги називають **відносними частотами** або **частості**. Загальна сума частостей дорівнює одиниці, тобто $\sum m/n = 1$, або $\sum m/n \cdot 100 = 100\%$, якщо частоти виражені у відсотках від n. Заміна частот частості не обов'язкова, але іноді вона буває корисною і навіть необхідною, тому що полегшує зіставлення одного варіаційного ряду з іншим, що особливо важливо в тих випадках, коли зіставляються ряди розрізняються за чисельністю складових їх варіант.

Сукупність числових значень ознаки розподіляється в безінтервальний або інтервальний варіаційний ряд у залежності від того, як варіює ознака – в широкому або вузькому діапазоні.

Для наочного вираження закономірності варіювання того чи іншого кількісної ознаки варіаційні ряди зображують у вигляді геометричних фігур у системі прямокутних координат. Так, якщо з'єднати прямими лініями геометричні точки, що зв'язують течії видів риб (відкладаються по осі абсцис) з їх частотами, що відкладаються по осі ординат, вийде лінійний графік, який має назву **варіаційної кривої** або **кривої розподілу**.

При побудові графіка безінтервального варіаційного ряду по осі абсцис відкладають значення видів риб, а по осі ординат – частоти. Висота перпендикулярів, підіймали по осі абсцис, відповідає частотах видів. Поєднуючи вершини перпендикулярів прямими лініями, одержують геометричну фігуру у вигляді багатокутника, який має назву **полігону розподілу частот** (рис. 2.1)

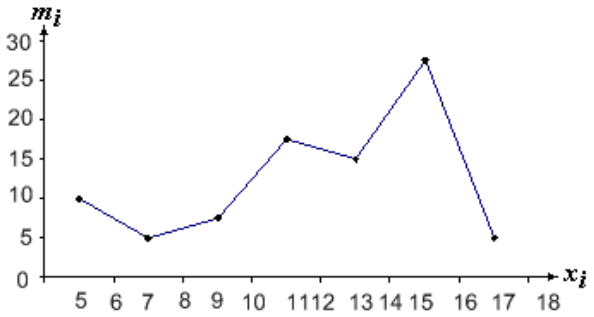


Рисунок 2.1 – Полігон розподілу частот

При побудові графіка інтервального варіаційного ряду по осі абсцис відкладають кордону видових інтервалів. У результаті виходить стовпчикова геометрична фігура, яка має назву **гістограма розподілу**

частот (рис. 2.2). Якщо з серединних точок вершин прямокутників гістограми опустити перпендикуляри на вісь абсцис, гістограма перетворюється на полігон розподілу. Поєднуючи точки вершин прямокутників гістограми прямими лініями, одержуємо варіаційну криву.

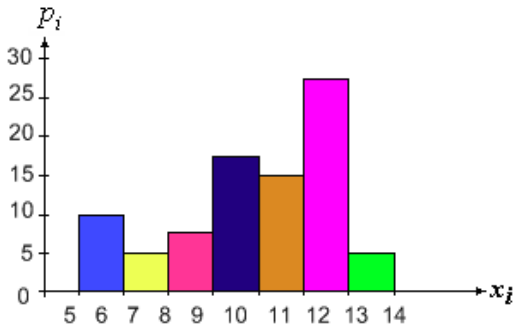


Рисунок 2.2 – Гістограма розподілу частот

Отже, будь-який варіаційний ряд у загальному вигляді описується графічно варіаційною кривою, яка має не тільки ілюстративне, але й аналітичне значення: вона служить свого роду відправним пунктом статистичного аналізу варіюють ознак.

Питання для самоперевірки:

1. Що називається генеральною сукупністю?
2. Що називається вибіркової сукупністю?
3. Що називається вибіркою?
4. Які бувають спостереження?
5. Що таке репрезентативність вибірки?
6. Що таке принцип рендомізації?
7. Як проводиться повторний відбір?
8. Як проводиться без повторний відбір?
9. Як відбувається групування первинних даних?
10. Як відбувається просте групування ознак?
11. Як відбувається складне групування ознак?
12. Що називається варіаційним рядом?
13. Що називається рядом розподілу?
14. Що називається варіаційною кривою або кривою розподілу?
15. Що таке полігон розподілу частот?
16. Що таке гістограма розподілу частот?

3 СЕРЕДНІ ВЕЛИЧИНИ І ПОКАЗНИКИ ВАРІАЦІЇ

3.1 Середні величини

Варіаційні ряди та їх графіки дають наочне уявлення про те, як варіює той чи інший кількісний ознака. Але вони недостатні для повної характеристики статистичної сукупності, оскільки містять багато деталей, охопити які без застосування зведених або узагальнюючих кількісних показників неможливо. Кількісні показники, які логічно і теоретично обгрунтовані і дозволяють судити про якісний своєрідності варіюють об'єктів і порівнювати їх між собою, називаються **статистичними характеристиками**. Найбільш важливі серед них середні величини та показники варіації ознак.

На відміну від індивідуальних числових характеристик середні величини мають більшу стійкість, здатність характеризувати групу однорідних одиниць одним (середнім) числом. Середній зріст, середня маса, продуктивність та інші середні – усе це поняття абстрактні про конкретні речі. Значення середніх полягає в їх властивості акумулювати або врівноважувати всі індивідуальні відхилення, в результаті чого виявляється то найбільш стійке й типове, що характеризує якісну своєрідність групового об'єкта, дозволяє відрізняти його від інших варіюють об'єктів.

За визначенням К. Гауса, істинної середньої служить така величина, сума квадратів відхилень від якої має найменшим значенням. Для біолога найбільший інтерес представляють статечні середні. Усі вони утворюються із загальної формули

$$M = \sqrt[k]{\frac{\sum x_i^k}{n}} \qquad \text{або} \qquad M = \left[\frac{\sum x_i^k}{n} \right]^{\frac{1}{k}}, \qquad (3.1)$$

де M – середня величина; x_i – варіанта (випадкова величина);

k – величина, що визначає вид середньої;

n – обсяг вибірки, на якому обчислюється середня;

$\sum$ – знак суми.

Так, при $k = 1$ виходить середня арифметична, при $k = -1$ – середня гармонійна і т.д.

Вибіркові середні, тобто величини, що характеризують сукупність вибіркових даних, прийнято позначати тими ж літерами, якими позначені варіанти, з тією лише різницею, що над буквою ставиться риса. Так, якщо ознаку позначити через X, то його числові значення – x, а середня арифметична – $\overline{x}$, середня гармонійна позначається символом $\overline{x}_h$, середня геометрична – символом $\overline{x}_g$ і т. д.

Крім степеневих середніх у біології застосовуються і структурно середні – *медіана, мода* та ін

Середні величини можуть характеризувати тільки однорідну масу варіант. Якщо середня отримана на неоднорідному в якісному відношенні матеріалі і вибрана неправильно, без урахування специфіки описуваного явища або процесу, вона виявиться фіктивною. При наявності різнорідних за складом даних їх необхідно групувати в окремі якісно однорідні групи і обчислювати групові або приватні середні.

Середні величини і показники варіації обчислюються як на групуватися, так і не групувати в варіаційний ряд матеріалі.

Існують три основні способи обчислення узагальнюючих характеристик:

1) основний, або спосіб добутків;

2) умовної середньої, або умовного нуля і довільного початку;

3) спосіб сум, заснований на кумуляції частот варіаційного ряду.

Спосіб сум в даний час застосовується рідко, тому розглядати його не будемо.

Основний спосіб, або спосіб добутків, заснований на використанні відхилень варіант даної сукупності від їх середньої величини.

Якщо варіанти у вибірці повторюються, то обчислення середньої величини і показників варіації значно спрощується, коли використовуються твори класових варіант на їх частоти (звідси і назва основного способу – «спосіб добутків»).

Спосіб умовної середньої. Обчислення статистичних характеристик варіаційного ряду основним способом трудомістко, особливо за наявності багатозначних чисел. Простіше розраховуються статистичні характеристики спрощеним **способом умовної середньої**. Сутність цього способу, заснованого на математичних властивості зазначених характеристик, полягає в наступному. Одну з варіант, чи будь-яке із значень класових варіант, умовно приймають за середню величину, позначивши її через A. Зазвичай в якості умовної середньої A береться варіанта або клас з найбільшою частотою (можна прийняти будь-яку варіанту або будь-який клас варіаційного ряду). Позначивши величину A, залишається знайти поправку, яку потрібно додати або відняти (дивлячись

по її знаку) від умовної середньої A, щоб отримати шукану величину середньої арифметичної. Ця поправка – **умовний момент першого порядку** – позначається символом b_1 і виражається формулою $b_1 = \sum p_i(x_i - A)/n$. Позначивши відхилення варіант від умовної середньої через a, отримаємо $b_1 = \sum pa/n$.

Звідси формула для визначення середньої арифметичної:

$$\overline{x} = A + \sum pa/n \ . \tag{3.2}$$

Дисперсія, яка визначається цим способом, виражається формулою

$$S_x^2 = \frac{n}{n-1}\left[\frac{\sum pa^2}{n} - \left(\frac{\sum pa}{n}\right)^2\right], \tag{3.3}$$

де $\sum pa^2 \Big/ n$ величина, що позначається символом b_2, називається **умовним моментом другого порядку**. Таким чином, дисперсія являє собою різницю між умовним моментом другого і квадратом умовного моменту першого порядку, помножену на $\frac{n}{n-1}-$ величину, яка називається **поправкою Бесселя**, тобто

$$S_x^2 = \left(b_2 - b_1^2\right)\frac{n}{n-1} \ , \tag{3.4}$$

де b_1 – умовний момент першого порядку;

b_2 – умовний момент другого порядку;

$\frac{n}{n-1}-$ поправкою Бесселя.

3.1.1 Середня арифметична та її властивості

Із загального сімейства статечних середніх найбільш часто використовується *середня арифметична* – одна з основних характеристик варіаційного ряду, що є центром розподілу, навколо якого групуються всі варіанти статистичної сукупності. Для негрупованих даних ця величина визначається як сума всіх членів сукупності, поділена на їх загальне число *n*. Так, якщо варіююча ознака позначити через *X*, то середня арифметична з

значенні цієї ознаки – $x_1, x_2, x_3, \ldots, x_n$, яка називається **простою**, виразиться наступною формулою:

$$\overline{x} = \frac{x_1 - x_2 - x_3 - \ldots - x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \qquad (3.5)$$

де x_i – значення варіант;

$\sum_{i=1}^{n}$ – знак підсумовування варіант у межах від першої до й варіанти;

n – загальне число варіант, що складають дану сукупність (обсяг вибірки).

Для даних, згрупованих з урахуванням повторюваності або ваги (p) окремих варіант, середня арифметична, яка називається **зваженою**, обчислюється за формулою

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i p_i \ , \qquad (3.6)$$

де p_i – частоти варіант або випадкових величин.

У тих випадках, коли доводиться поєднувати характеристики кількох однорідних вибірок, середня $\overline{\overline{x}}$ з суми групових чи часних $\overline{x}_1, \overline{x}_2, \overline{x}_3, \ldots, \overline{x}_k$ середніх, вагами яких є обсяги груп $n_1, n_2, n_3, \ldots, n_k$ визначають за формулою

$$\overline{\overline{x}} = \frac{\overline{x}_1 n_1 + \overline{x}_2 n_2 + \overline{x}_3 n_3 + \ldots + \overline{x}_k n_k}{n_1 + n_2 + n_3 + \ldots + n_k} = \frac{\sum(x_i n_i)}{\sum n_i}. \qquad (3.7)$$

Середня арифметична має ряд важливих властивостей:

1. якщо кожну варіанту сукупності зменшити або збільшити на якесь довільне додатне число A, то і середня зменшиться або збільшиться на стільки ж.

Доказ:

$$\overline{x}^* = \frac{\sum(xi - A)pi}{\sum p_i} = \frac{\sum x_i p_i}{\sum p_i} - \frac{A\sum p_i}{\sum p_i} = \overline{x} - A \ . \qquad (3.8)$$

Звідси випливає, що середню $\overline{x}$ можна обчислити за зменшеним на A варіантам ряду, додавши до отриманого значення вираховане з варіант число A, тобто $\overline{x} = \overline{x}^* + A$.

2. Якщо кожну варіанту розділити або помножити на одне і те ж число A, то й середня арифметична зміниться в стільки ж разів.

Доказ:

$$\overline{x}^* = \frac{\sum\left(\frac{x_i}{A}\right)p_i}{\sum p_i} = \frac{\sum x_i p_i}{A\sum p_i} = \frac{\overline{x}}{A} \ . \qquad (3.9)$$

Ця властивість дозволяє обчислювати середню ($\overline{x}$) спрощеному способом, зменшивши попередньо всі варіанти ряду в A раз, а потім помноживши отриманий результат на A, тобто $\overline{x} = \overline{x}^* \cdot A$.

3. Сума добутків відхилень варіант від їх середньої арифметичної на відповідні їм частоти дорівнює нулю.

Доказ:

$$\sum p_i(x_i - \overline{x}) = \sum p_i x_i - \sum p_i \overline{x} = \sum p_i\left(\frac{\sum p_i x_i}{\sum p_i} - \overline{x}\right) = \sum p_i(\overline{x} - \overline{x}) = 0 \ . \quad (3.10)$$

Ця властивість дозволяє в кожному конкретному випадку перевірити правильність обчислення середньої арифметичної, що є центром розподілу.

4. Сума квадратів відхилень варіант від їх середньої $\overline{x}$ менше суми квадратів відхилень тих же варіант від будь-якої іншої величини A, що не дорівнює $\overline{x}$:

$$\sum\left(x_i - \overline{x}\right)^2 < \sum\left(xi - A\right)^2 . \qquad (3.11)$$

Ця властивість дозволяє визначати середню арифметичну не прямим, а непрямим способом – за допомогою числа A, званого умовної середньої.

3.2 Показники варіації

3.2.1 Ліміти, розмах варіації, дисперсія та її властивості, середньоквадратичне відхилення, коефіцієнт варіації

Середні величини не містять повної інформації про варіюють об'єктах. При однакових середніх характеризуються ними ознаки можуть

відрізнятися за величиною варіації. Тому поряд із середньою величиною для більш повної характеристики варіаційного ряду повинні обчислюватися і показники варіації. Одним з таких показників служать **ліміти**, що позначаються символом *lim* (від лат. *Limes* – межа). В біологічній статистики під цим терміном розуміються значення мінімальної ($x_{\min}$) і максимальної ($x_{\max}$) варіант, між якими розподіляються всі члени даної сукупності.

Розмах варіації – інший показник, що характеризує варіювання ознак. Він визначається по різниці максимальної та мінімальної варіант даної сукупності: $R = x_{\max} - x_{\min}$.

Дисперсія та її властивості. Ліміти і розмах варіації конкретні і прості, в чому і полягає їх позитивна сторона як показників варіації. Але вони здатні сильно змінюватися при повторних вибірках з однієї і тієї ж генеральної сукупності. Крім того, вони не відображають суттєві риси варіювання, що можна показати на наступному прикладі. Візьмемо два ряди розподілу, варіанти яких мають один і той же вага, що дорівнює одиниці:

x_1 . . .	10	15	20	25	30	35	40	45	50	$x_1 =$	30
x_2 . . .	10	28	28	30	30	30	32	32	50	$x_2 =$	30

За кількістю варіант, лімітами і розмаху варіації ці ряди не відрізняються один від одного; їх середні арифметичні також рівні між собою, але характер варіювання у них різний, проте це не відбилося на величині лімітів і розмах варіації.

Цього недоліку позбавлений середній квадрат відхилень варіант даної сукупності від їх середньої величини – показник, кий називається **середнім квадратом (відхилень), дисперсією** (від лат. *Dispersio* – розсіювання). Дисперсія – найважливіша характеристика варіаційного ряду. Дисперсія генеральної сукупності позначається символом σ^2, а дисперсія вибірки – S^2 і визначається за формулою

$$S_x^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 p_i}{n-1}, \qquad (3.12)$$

де $\sum_{i=1}^{k}$ – знак підсумовування добутків відхилень варіант x_i, від їх середньої $\bar{x}$ на ваги або частоти p_i цих відхилень в межах від першого до k–го класу; n – загальне число спостережень, або обсяг вибірки.

Приписана до символу дисперсії буква x позначає, що цей показник у даному випадку характеризує варіювання числових значень x_i ознаки X навколо їхньої середньої $\overline{x}$. Величина $n-1$ – число вільно варіюють одиниць або елементів у складі даної чисельно обмеженої сукупності – називається **числом степенів вільності**. Так, якщо сукупність складається з n ряду членів – $x_1, x_2, x_3, \ldots, x_n$, тобто має об'єм, рівний n, і характеризується середньою величиною $\overline{x}$, то будь-який член цієї сукупності може мати яке завгодно значення, не змінюючи при цьому середню $\overline{x}$, крім однієї варіанти, значення якої визначається різницею між сумою значень всіх інших варіант і значенням $n\overline{x}$. Отже, одна варіанту чисельно обмеженою сукупності не має свободи варіації і число степенів вільності буде дорівнює обсягу вибірки без одиниці, тобто $n-1$. При наявності не одного, а кількох обмежень свободи варіації число степенів вільності $k = n - \nu$, $n-$де обсяг вибірки, $\nu-$число обмежень свободи варіації. В інших випадках, застосовуючи замість n число степенів вільності k, отримуємо **незсунену дисперсію**, яка є більш точною оцінкою генерального параметра.

Цінність дисперсії полягає в тому, що, будучи мірою варіювання числових значень ознаки навколо їхньої середньої арифметичної, вона вимірює і внутрішню мінливість значень ознаки, що залежить від різниць між спостереженнями. Перевага дисперсії перед іншими показниками варіації полягає також і в тому, що вона розкладається на складові компоненти, дозволяючи тим самим оцінювати вплив різних чинників на величину обліковується ознаки.

Як і середня арифметична, **дисперсія має ряд важливих властивостей:**

1. Якщо кожну варіанту сукупності зменшити або збільшити на одне і те ж постійне число A, то дисперсія не зміниться.

Доказ:

$$S_x^2 = \frac{1}{n-1}\sum\left[(x_i - A) - (\overline{x} - A)\right]^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2 \ . \qquad (3.13)$$

З цієї властивості випливає, що дисперсію можна обчислити не тільки за значеннями ознаки, але і за їх відхилень від якої-небудь постійної величини A.

2. Якщо кожну варіанту розділити (або помножити) на одне і те ж постійне число A, то дисперсія зменшиться (або збільшиться) в A^2 раз.

Доказ:

$$S_x^2 = \frac{1}{n-1}\sum\left(\frac{x_i}{A} - \frac{\bar{x}}{A}\right)^2 = \frac{1}{A^2-(n-1)}\sum(x_i - \bar{x})^2 \qquad (3.14)$$

або

$$S_x^2 = \frac{1}{n-1}\sum\left(x_i A - \bar{x}A\right)^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 A^2 \ . \qquad (3.15)$$

З цієї властивості випливає, що за наявності в сукупності багатозначних варіант їх можна скоротити на якесь постійне число A і за результатами обчислити дисперсію, а потім помножити отриману величину на квадрат загального дільника A^2, що дасть у результаті шукану величину дисперсії.

Слід також мати на увазі, сто замість $\sum(x_i - \bar{x})^2$ можна використовувати:

$$\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}; \quad n\left[\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2\right]; \quad \sum x_i^2 - n\bar{x}^2.$$

Звідси виходять такі робочі формули, зручні при обчисленні дисперсії безпосередньо за значеннями варіюючої ознаки:

$$S_x^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right], \qquad (3.16)$$

$$S_x^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\cdot\bar{x}^2\right), \qquad (3.17)$$

$$S_x^2 = \frac{n}{n-1}\left[\frac{\sum x_i^2}{n}\left(-\frac{\sum x_i}{n}\right)^2\right]. \qquad (3.18)$$

Середнє квадратичне відхилення. Поряд з дисперсією найважливішою характеристикою варіації служить середнє квадратичне відхилення S_x, що представляє корінь квадратний з дисперсії:

$$S_x = \sqrt{\frac{\sum_{i=1}^{k} p_i (x_i - \overline{x})^2}{n-1}} \ . \qquad (3.19)$$

Ця величина, яка називається також **стандартним відхиленням**, виражається в тих же одиницях, що і варіанти сукупності, і часто виявляється більш зручною характеристикою варіювання, ніж дисперсія. Чим сильніше варіює ознака, тим більше величина цього показника, і, навпаки, чим слабкіший варіює ознака, тим менше середнє квадратичне відхилення.

Дисперсія і середнє квадратичне відхилення характеризують не лише величину, а й специфіку варіювання, а також застосовні і для порівняльної оцінки однойменних середніх величин. У практиці ж досить часто доводиться порівнювати мінливість ознак, виражених різними одиницями. У таких випадках використовують не абсолютні, а відносні показники варіації. Дисперсія і середнє квадратичне відхилення як величини, які виражаються тими ж одиницями, що і характеризується ними ознака, для оцінки мінливості різнойменних величин непридатні. Одним з відносних показників варіації є **коефіцієнт варіації**. Цей показник являє собою середнє квадратичне відхилення, виражене у відсотках від величини середньої арифметичної:

$$V = \frac{S_x}{\overline{x}} \cdot 100\% \ , \qquad (3.20)$$

де V – коефіцієнт варіації;
S_x – середнє квадратичне відхилення;
$\overline{x}$ – середня арифметична.

3.2.2 Нормоване відхиленням

У сфері статистичного аналізу важливе місце займає **нормоване відхилення** – показник, що позначається буквою t і представляє відхилення тієї чи іншої варіанти від середньої величини, віднесене до величини середнього квадратичного відхилення:

$$t = (x_i - \overline{x}) / S_x \ . \qquad (3.21)$$

Цей показник дозволяє встановлювати, на скільки «сигм» окремі члени даної сукупності відхиляються від середнього рівня обліковується ознаки.

Середня гармонійна. У деяких випадках при обчисленні середньої величини використовують не абсолютні значення варіюючої ознаки, а зворотні числа окремих варіант. Отримана при цьому характеристика називається **середньою гармонійною** і позначається символом $\overline{x}_h$. Середня гармонійна, як і інші середні, може бути простою і зваженою. представляє відношення обсягу вибірки (n) до суми **Проста середня гармонійна** зворотних значень ознаки:

$$\overline{x}_h = n : \sum_{i=1}^{n} \left(\frac{1}{x_i} \right), \tag{3.22}$$

зважена середня гармонійна виражається наступною формулою:

$$\overline{x}_h = n : \sum_{i=1}^{k} \left(\frac{1}{x_i} \cdot p_i \right), \tag{3.23}$$

де $1/x_i$ – зворотні значення варіант, а p_i – їх ваги (частоти);

n – число варіант, з яких розраховується середня $\overline{x}_h$, тобто обсяг вибірки.

Середня квадратична. При вираженні ознак заходами площі більш точною характеристикою буде середня квадратична, що позначається символом $\overline{x}_q$. Ця величина дорівнює кореню квадратному із суми квадратів варіант, віднесеної до їх загального числа в даній вибірці, тобто

$$\overline{x}_q = \sqrt{\frac{\sum x_i^2}{n}}, \tag{3.24}$$

або при повторюваності окремих варіант

$$\overline{x}_q = \sqrt{\frac{\sum p_i x_i^2}{n}}. \tag{3.25}$$

Середня кубічна – більш точна характеристика об'ємних ознак. Вона позначається символом $\overline{x}_k$ і дорівнює кореню кубічному з суми кубів варіант, поділеній на їх кількість, тобто

$$\overline{x}_k = \sqrt[3]{\frac{\sum x_i^3}{n}} \ , \qquad (3.26)$$

або з урахуванням повторюваності окремих варіант

$$\overline{x}_k = \sqrt[3]{\frac{\sum p_i x_i^3}{n_i}} \ . \qquad (3.27)$$

Середня геометрична – більш точна характеристика при визначенні середніх збільшень або при збільшенні лінійних розмірів тіла, приросту чисельності популяції за певний проміжок часу. Вона позначається символом $\overline{x}_g$ і дорівнює кореню n-го ступеня з добутків членів ряду:

$$\overline{x}_g = \sqrt[n]{x_1 x_2 x_3 \ldots x_n} \ . \qquad (3.28)$$

Зазвичай середня геометрична визначається за допомогою десяткових логарифмів за формулою

$$\lg \overline{x}_g = \frac{1}{n-1}(\lg x_1 + \lg x_2 + \ldots + \lg x_n) = \frac{1}{n-1}\sum_{i=1}^{n} \lg x_i \ , \qquad (3.29)$$

тобто логарифм середньої геометричної дорівнює середній арифметичній із логарифмів всіх членів ряду. При цьому відхилення логарифмів окремих варіант ($\lg x_i$) від логарифма середньої геометричної у сумі дорівнюють нулю, тобто $\sum(\lg x_i - \lg \overline{x}_g) = 0$ (Основна властивість середніх величин).

3.3 Структурні середні

Медіана емпіричного розподілу – середня, щодо якої ряд розподілу ділиться на дві половини: в обидві сторони від медіани розташовується однакове число членів ряду (варіант). Медіана позначається символом *Ì å* (від лат. *Mediana* – середня). На невеликих вибірках визначити медіану досить легко. Для цього сукупність спостережень ранжують по зростаючим значенням ознаки, і якщо число членів ряду непарне, то центральна варіанти і буде його медіаною. При парному числі членів ряду

медіана визначається за напівсумі двох сусідніх варіант, розташованих у центрі рада.

Якщо вибірка розподілена в варіаційний ряд, медіана визначається наступним чином:

$$Me = x_e + \frac{c\left(\frac{n}{2} - m^*\right)}{m_e}, \tag{3.30}$$

де

n – об'єм вибірки (за умови непарного значення об'єму береться $n+1$);

m^* – накопичена частота до медіанного інтервалу;

x_e – початок медіанного інтервалу;

c – довжина медіанного інтервалу;

m_e – частота медіанного інтервалу.

Мода ($\grave{I}\ \hat{i}$) – величина, яка зустрічається в даній сукупності найбільш часто. Градація (інтервал) з найбільшою частотою називається **модальним**.

Якщо ряд є *згрупованим* та *ранжування* проводилося *в бік зростання* значень випадкової величини, то моду визначають за такою формулою:

$$M_o = x_0 + c\frac{\left(m_i - m_{i-1}\right)}{\left(2m_i - m_{i-1} - m_{i+1}\right)}, \tag{3.31}$$

де x_0 – початок модального інтервалу;

c – довжина модального інтервалу;

m_i – емпірична частота модального інтервалу;

m_{i-1}, m_{i+1} – частоти попереднього і наступного за модальним часткових інтервалів.

Питання для самоперевірки:

1. Що називається статистичними характеристиками?
2. Що представляє собою середня величина?
3. Що називається основним способом або способом добутків?
4. Що називається способом умовної середньої?
5. Охарактеризуйте умовний момент першого порядку.
6. Охарактеризуйте умовний момент другого порядку.
7. Що таке середня арифметична і які її основні властивості?
8. Що називається простою середньою арифметичною?
9. Перелічіть основні показники варіації.
10. Що таке ліміти?
11. Що таке розмах варіації?
12. Що таке дисперсія і які її основні властивості?
13. Що таке середній квадрат відхилень?
14. Що називається середньоквадратичним відхиленням?
15. Що називається коефіцієнтом варіації?
16. Охарактеризуйте стандартне відхилення випадкової величини.
17. Що таке нормоване відхилення?
18. Яка величина називається середньою гармонічною?
19. Що таке проста та зважена середньо гармонічна величина?
20. Що таке середня кубічна величина?
21. Що таке середня геометрична величина?
22. Що таке медіана?
23. Що називається початком медіанного інтервалу?
24. Що називається об'ємом вибірки випадкових величин?
25. Що таке довжина медіанного інтервалу?
26. Дайте визначення накопиченої частоті і частоті медіанного інтервалу.
27. Що таке мода?
28. Що називається початком модального інтервалу?
29. Що таке довжина модального інтервалу?
30. Дайте визначення частотам модального інтервалу.

4 СТАТИСТИЧНІ ОЦІНКИ ГЕНЕРАЛЬНИХ ПАРАМЕТРІВ

4.1 Точкові оцінки

Під терміном **«оцінка»** розуміється не тільки той чи інший вибірковий показник, а й сам процес оцінювання генеральних параметрів – встановлення значущості або статистичної достовірності вибіркових оцінок.

Вибіркові характеристики – середня арифметична ($\overline{x}$), середнє квадратичне відхилення (S_x) та інші – величини випадкові, що варіюють навколо своїх генеральних параметрів – генеральної середньої (μ) і дисперсії (σ_x^2) або стандартного відхилення (σ_x).

Вибіркові характеристики розглядаються як наближені значення або точкові оцінки відповідних генеральних параметрів, які, як правило, залишаються невідомими. Вибіркова середня ($\overline{x}$) служить оцінкою генеральної середньої (μ), вибіркова дисперсія (S_x^2) є оцінкою дисперсії генеральної (σ_x^2) сукупності, а середнє квадратичне відхилення (S_x) розглядається в якості точкової оцінки стандартного відхилення – (σ_x) генеральної сукупності, яка вважається як сукупність не обмежено великого обсягу $(N \to \infty)$.

Вимоги, що пред'являються до оцінок. Щоб вибіркові характеристики мали достатньо добре наближення до генеральних параметрами, вони повинні задовольняти вимогам умотивованості, незсуненості та ефективності. Точкові оцінки називаються **умотивованими**, якщо при збільшенні числа випробувань $(n \to \infty)$ вони прагнуть до величини оцінюваних параметрів. Для математичного очікування $\mu(X)$ умотивованою оцінкою є вибіркова середня ($\overline{x}$), а для генеральної дисперсії (σ_x^2) умотивованої оцінкою служить вибіркова дисперсія (S_x^2). Оцінка називається **незсуненою**, якщо вона не містить систематичної помилки і її математичне сподівання збігається з генеральним параметром. Вибіркова середня ($\overline{x}$) – незсунена оцінка генеральної середньої $\mu(X)$, тоді як вибіркова дисперсія (S_x^2) і середнє квадратичне відхилення (S_x) виявляються оцінками, зміщеними щодо параметрів σ_x^2 та σ_x. Згадаймо одна з властивостей середньої: для кожної з можливих вибірок сума квадратів відхилень варіант (x_i) від середньої

($\bar{x}$) буде менше, ніж сума квадратів відхилень тих же варіант (x_i) від будь-якої іншої величини A, тобто $\sum(x_1-\bar{x})^2 < \sum(x_1-A)^2$, не виключаючи і генеральну середню $\mu(X)$, тобто $\sum(x_1-\bar{x})^2 < \sum(x_1-\mu(X))^2$. Звідси випливає, що при визначенні вибіркової дисперсії як середньої суми квадратів відхилень, тобто $S_x^2 = \frac{\sum(x_1-\bar{x})^2}{n}$, отримуємо занижену або зміщену оцінку генерального параметра σ_x^2 на величину рівну $\frac{n}{n-1}$. Щоб отримати незміщене оцінку генеральної дисперсії, необхідно вибіркову дисперсію виправити, помноживши її на величину $\frac{n}{n-1}$, тобто $S_x^2 = \frac{\sum(x-\bar{x})^2}{n} \cdot \frac{n}{n-1} = \frac{\sum(x_i-\bar{x})^2}{n-1}$.

Іншими словами, при обчисленні вибіркової дисперсії та середнього квадратичного відхилення суму квадратів відхилень потрібно відносити не до числа спостережень (n), а до числа степенів вільності ($n-1$), чим і усувається систематична помилка, що виникає при обчисленні вибіркової дисперсії як середнього квадрата відхилення варіант (x_i) від їх середньої ($\bar{x}$).

Незсунена оцінка називається **ефективною** чи **найкращою**, якщо вона має найменшу дисперсію порівняно з іншими оцінками одного і того ж генерального параметра.

Помилки вибіркових показників

Як правило, вибіркові характеристики не збігаються за абсолютною величиною з відповідними генеральними параметрами. Величина відхилення вибіркового показника від його генерального параметра називається **статистичної помилкою** цього показника або **помилкою репрезентативності**. **Статистичні помилки** – це не помилки, які допускаються при вимірюванні біологічних об'єктів. Вони виникають виключно в процесі відбору варіант з генеральної сукупності і до помилок вимірів відношення не мають.

Величина помилки репрезентативності вимірюється середнім квадратичним відхиленням, яке є не лише характеристикою варіювання тієї чи іншої ознаки, але і служить мірою «помилки» окремих варіант, якщо вони використовуються в якості оцінки генеральних параметрів.

З математичної статистики відомо, що вибіркові середні $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_k$, обчислені на матеріалі незалежних вибірок з однієї і тієї ж генеральної сукупності, що нормально розподіляється, варіюють навколо

генерального параметра (μ) в $\sqrt{n}$ разів менше, ніж окремі варіанти даної сукупності. Звідси помилка репрезентативності вибіркової середньої виражається формулою

$$S_{\overline{x}} = \frac{S_x}{\sqrt{n}} = \sqrt{\frac{S_x^2}{n}} \ . \tag{4.1}$$

Статистичну помилку вибіркової середньої можна висловити і у вигляді такої формули, яка іноді зручніше в роботі:

$$S_{\overline{x}} = \sqrt{\frac{\sum\left(x_i - \overline{x}\right)^2}{n\left(n-1\right)}} \ , \tag{4.2}$$

Зручними є й інші формули, в які трансформуються основні формули помилки:

$$S_{\overline{x}} = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n\left(n-1\right)}} \ , \tag{4.3}$$

або

$$S_{\overline{x}} = \sqrt{\frac{1}{n-1}\left[\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2\right]} \ . \tag{4.4}$$

Коли середня арифметична ($\overline{x}$) обчислюється способом умовної середньої (A), помилка визначається за наступною формулою

$$S_{\overline{x}} = \sqrt{\frac{1}{n-1}\left[\frac{\sum pa^2}{n} - \left(\frac{\sum pa}{n}\right)^2\right]} \ , \tag{4.5}$$

де $a = \left(x_i - A\right)$ – відхилення варіант $\left(x_i\right)$ від умовної A;

p – повторюваність або частота відхилення

Коли виникає необхідність знайти середню з декількох приватних середніх з їх помилками, помилка загальної середньої обчислюється за формулою

$$S_x = \frac{1}{K}\sqrt{S\frac{2}{x_1} + S\frac{2}{x_2} + \cdots + S\frac{2}{x_k}} \ . \qquad (4.6)$$

Помилка твори двох вибіркових середніх з їх помилками визначається за формулою

$$S_{\bar{x}_1\bar{x}_2} = \bar{x}_1\bar{x}_2\sqrt{\left(\frac{S_{\bar{x}_1}}{x_1}\right)^2 + \left(\frac{S_{\bar{x}_2}}{x_2}\right)^2} \ . \qquad (4.7)$$

Помилка частки від розподілу вибіркових середніх з їх помилками обчислюється за формулою

$$S_{\bar{x}_1/\bar{x}_2} = \frac{\bar{x}_1}{\bar{x}_2}\sqrt{\left(\frac{S_{\bar{x}_1}}{x_1}\right)^2 + \left(\frac{S_{\bar{x}_2}}{x_2}\right)^2} \ . \qquad (4.8)$$

Статистичні помилки характеризують варіювання вибіркових показників навколо їх генеральних параметрів. Вони мають ті ж властивості, що й середнє квадратичне відхилення. Лише одна властивість специфічно для помилок репрезентативності: вони зменшуються при збільшенні обсягу вибірки, тобто при $n \to \infty$, $S_{\bar{x}} \to 0$. Це властивість статистичних помилок обумовлено дією закону великих чисел, по якому найбільш ймовірний результат виходить при найбільшій кількості випробувань. Звідси зрозуміло значення помилки: вона вказує на точність, з якою вибіркові показники репрезентують генеральні параметри. Чим менше помилка, тим ближче вибіркова характеристика до величини генерального параметра, і, навпаки, чим більше помилка, тим менш точно репрезентує вибіркова характеристика її генеральний параметр. Отже, за властивістю статистичної помилки, яка при $n \to \infty$ прагне до нуля, можна судити про спроможність оцінок.

Показник точності визначення середньої. Про близькість вибіркової середньої до генерального параметру можна судити по відношенню помилки репрезентативності до супроводжуваної нею середню величину. Цей показник C_S визначається за формулою

$$C_S = \frac{S_{\bar{x}}}{\bar{x}} \cdot 100\% \ . \qquad (4.9)$$

Показник точності можна обчислити і за такою формулою:

$$C_S = \frac{\upsilon\%}{\sqrt{n}}\ , \tag{4.10}$$

де $\upsilon\%$ – коефіцієнт варіації, а n – обсяг вибірки.

Показник C_S знайшов широке застосування особливо в досвідченої агрономії, де його вважають корисною характеристикою при оцінці результатів дослідів. Точність середніх показників, якими оцінюють результати спостережень, вважається цілком задовільною, якщо коефіцієнт C_S не перевищує 3–5%.

Точкові оцінки при відомому обсязі генеральної сукупності.

На величині помилки вибіркових показників позначаються і способи відбору варіант з генеральної сукупності і ступінь варіювання ознаки. Чим сильніше варіює ознака, тим більше при інших рівних умовах буде помилка вибіркових показників, і, навпаки, при слабкому варіюванні ознаки помилка вибіркових показників буде менше. При безповторному відборі варіант з чисельно обмеженою генеральної сукупності, особливо в тих випадках, коли вибірка досить значна (при $n \geq 25\%$ від N), помилка вибіркової середньої, яка обчислюється за формулою (4.1), виявляється неточною, дещо завищеною. Враховуючи цю обставину, К. Пірсон (1899) запропонував поправочний коефіцієнт, що дорівнює $\sqrt{\frac{N-n}{N-1}}$ на який варто множити помилку середньої, що характеризує безповторну вибірку з генеральної сукупності, що нормально розподіляється. Формула помилки вибіркової середньої в таких випадках набуває наступний вигляд:

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{S_x^2}{n}\left(\frac{N-n}{N-1}\right)}\ . \tag{4.11}$$

Коефіцієнт $\frac{N-n}{N-1}$ можна замінити на наближену величину $1-\frac{n}{N}$, де ставлення $\frac{n}{N}$ є частка вибірки. Чим більше ця частка, тим сильніше позначиться поправка на величині помилки. Якщо ж частка мала, що трапляється часто, поправка близька до одиниці і величина помилки практично не зміниться.

Щоб визначити помилку цієї величини, необхідно спочатку знайти зважену дисперсію ($\overline{S}_x^2$) для даного комплексу. Ця величина визначається за формулою

$$\overline{S}_x^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \ldots + (n_k-1)S_k^2}{N-v} = \frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{N-v}\ , \quad (4.12)$$

де $n_i - 1$ – число степенів вільності для вибіркових груп, що входять у статистичний комплекс, на яке зважуються групові дисперсії (S_i^2);

$N = \sum n_i$ – загальна кількість спостережень або обсяг комплексу,

v – число обмежень свободи варіації, рівне чисельності груп, що входять до складу даного комплексу.

4.2 Інтервальні оцінки

Довірчий інтервал для генеральної середньої. За відомим значенням вибіркових характеристик можна встановити інтервал, в якому з тією чи іншою ймовірністю знаходиться величина генерального параметра. Вірогідність, визнані достатніми для впевнених суджень про генеральних параметрах на основі вибіркових показників, називаються **довірчими**. Звичайно як довірчих використовують такі рівні або пороги ймовірності: P_1=0,95; P_2=0,99; P_3=0,999. Це означає, що при оцінюванні генеральних параметрів за вибірковими характеристиками ми ризикуємо помилитися у першому випадку один раз на 20 випробувань, в другому випадку – один раз на 100 випробувань і в третьому випадку – один раз на 1000 випробувань.

Вибір того чи іншого рівня ймовірності здійснюється дослідником виходячи з практичних міркувань і тієї відповідальності, з якою робляться висновки про генеральних параметрах за результатами вибіркових спостережень.

Вірогідність відхилення будь варіанти (x_i) сукупності, що нормально розподіляється, від центру розподілу (μ) визначається функцією нормованого відхилення (t), і довірчий інтервал для невідомого параметра буде наступний:

$$-t \le \frac{x_i - \mu}{S_x} \le +t \quad (4.13)$$

або

$$x_i - t\,S_x \le \mu \le x_i + t\,S_x \ . \tag{4.14}$$

Оскільки вибіркові середні $(\overline{x})$ варіюють навколо генеральної середньої μ в $\sqrt{n}$ разів менше, ніж окремо розвинені варіанти, границі довірчого інтервалу для генеральної середньої μ, що встановлюються за величиною вибіркової середньої, будуть наступні:

$$\overline{x} - t\frac{S_x}{\sqrt{n}} \le \mu \le \overline{x} + t\frac{S_x}{\sqrt{n}} \ . \tag{4.15}$$

Маючи на увазі, що $\frac{S_x}{\sqrt{n}} = S_{\overline{x}}$, довірчий інтервал для генеральної середньої (μ) можна виразити так $\overline{x} - t\,S_{\overline{x}} \le \mu \le \overline{x} + t\,S_{\overline{x}}$

Тут t – нормоване відхилення, яке визначається рівнем імовірності. У компактній формі границі довірчого інтервалу можна записати так: $\overline{x} \pm t\,S_{\overline{x}}$.

Довірчий інтервал для частки. Довірчий інтервал для генеральної частки (P) встановлюється таким же способом, як і для генеральної середньої, тобто

$$p - t\,S_{pf} \le P \le p + t\,S_{pf} \ . \tag{4.16}$$

Тут P – генеральна, а $p-$ вибіркова частки; $t-$ нормоване відхилення чи критерій, за яким з певною ймовірністю встановлюються границі довірчого інтервалу; $S_{pf}-$ помилка вибіркової частки, яка визначається за формулою:

$$S_{pf} = \sqrt{\frac{p(1-p)}{n}} \ . \tag{4.17}$$

Якщо замість частки в якість оцінки генерального параметра використовується відсоткова частота, її помилка обчислюється за наступною формулою:

$$S_{pf} = \sqrt{\frac{p(100-p)}{n}} \ . \tag{4.18}$$

Границі $p \pm t S_{pf}$ довірчого інтервалу для генеральної частки встановлюються з достатньою точністю в тих випадках, коли вибіркові частки рівні або не сильно відхиляються від 50% чисельності груп. Якщо ж вибіркові частки не рівні (75% < p <25%) і тим більш близькі до нуля і одиниці, довірчі границі для генеральної частки слід визначати за допомогою допоміжної величини φ:

$$\varphi = 2 \arcsin \sqrt{p} \ . \qquad (4.19)$$

Ця величина, запропонована Р. Е. Фішером, має розподіл, близький до нормального. Її параметром служить вибіркова помилка, рівна $1/\sqrt{n}$. Значення φ залежать тільки від p. Для практичного використання цієї величини складена спеціальна таблиця, в якій містяться значення φ для різних значень частки p, вираженої у відсотках, тобто для значень $p = m/n \cdot 100\%$. Застосовуючи величину, вибіркові частки p_1 і p_2, коригують введенням поправки Йейтса, рівною $(1/2)n$. Вона вираховується з більшою і додається до меншій частці.

4.3 Статистичні характеристики при альтернативному угрупованню випадкових величин (варіант)

Групування вибіркових даних на дві протиставлювані один одному групи називається **альтернативною**.

Чисельність альтернатив виражається в частках одиниці, а також у відсотках від їх загальної кількості (n). Позначивши частку варіант, що володіють даною ознакою, через p, отримаємо $p = m/n$. Тоді інша частка варіант тієї ж сукупності, у яких дана ознака відсутня, що позначається буквою q, виразиться відношенням $q = (n - m)/n = 1 - m/n = 1 - p$. Щоб висловити чисельність груп у відсотках від їх загальної кількості, досить кожну частку помножити на 100:

$$p = m/n \cdot 100\% \ , \qquad (4.20)$$

$$q = \frac{n - m}{n} \cdot 100\% = 100 - p \ , \qquad (4.21)$$

Очевидно, $p + q = 1$ і $p\% + q\% = 100$.

Відносні частоти, або частки, при альтернативній угрупованню варіант виконують таку ж роль, яку виконують середні величини для рядової мінливості ознак, коли вибірки розподіляються в варіаційний ряд.

Характеристикою варіювання протиставлюваних один одному груп, як і при рядовій мінливості ознак, служить середнє квадратичне відхилення (S_p). Коли альтернативи виражені в частках одиниці, цей показник визначається за формулою

$$S_p = \sqrt{p(1-p)} = \sqrt{pq} \ . \qquad (4.22)$$

Якщо ж альтернативи виражені у відсотках від їх загальної кількості, формула (4.22) набуває такий вигляд:

$$S_p = \sqrt{p\%(100-p\%)} \ , \qquad (4.23)$$

І нарешті, у тих випадках, коли альтернативи виражаються абсолютними числами, середнє квадратичне відхилення визначається за формулою:

$$S_p = \sqrt{npq} \ , \qquad (4.24)$$

де n – чисельність альтернатив виражається в частках одиниці, а також у відсотках від їх загальної кількості;

p – частка варіант, що володіють даними ознакою;

q – частка варіант тієї ж сукупності, у яких дана ознака відсутня.

Може бути квадратичне відхилення в рівній мірі характеризує варіювання обох протиставлюваних один одному груп.

Питання для самоперевірки:

1. Що називається статистичними оцінками генеральних параметрів?
2. Що називається оцінками параметрів, і які вони бувають?
3. Дайте визначення точковим оцінками параметрів.
4. Які вимоги пред'являються до точкових оцінками параметрів?
5. Що таке помилки вибіркових показників?
6. Що називається статистичною помилкою?
7. Що називається помилкою репрезентативності?
8. Дайте визначення показника точності визначення середньої.
9. Що таке інтервальні оцінки?
10. Дайте визначення довірчого інтервалу для частки.
11. Що таке альтернативна групування випадкових величин?

5 ЗАКОНИ РОЗПОДІЛУ

Законом розподілу випадкової величини називають будь-яку відповідність між можливими значеннями випадкової величини та їх ймовірностями. Для випадкової величини закон розподілу є вичерпною характеристикою. Знання закону, якому підпорядковується та чи інша біологічна величина, дає можливість методично правильно організувати дослідження статистичної структури цих величин. Підібравши закон розподілу до статистичного ряду (вібірки), можна розрахувати ймовірність того, що випадкова величина знаходиться у заданому інтервалі або ймовірність того, що випадкова величина прийми значення менше (більше) деякого конкретного числа із області значення цієї випадкової величини.

Випадкові події. Результат, або результат окремого випробування, називається **подією**. Реалізація того чи іншого значення варіюючої ознаки є випадкове подія. Під випробуванням ж розуміється певний комплекс умов, необхідних для того, щоб міг здійснитися той чи інший результат, тобто відбутися або не відбутися очікувана подія. Події A, B, C називаються **несумісними**, якщо в умовах випробування кожен раз можливо поява тільки одного з них. Якщо ж в даних умовах поява події A не виключає появи іншої події – B або C, вони називаються **сумісними**. Події A і $\overline{A}$ (тобто не $\overline{A}$) називаються **протилежними**, якщо в умовах випробування вони єдино можливі і несумісні. Це події єдино можливі, несумісні і протилежні. Здійснення одного з них залежить від багатьох випадкових причин, повністю врахувати які неможливо. Передбачити появу випадкової події в окремих випробуваннях можна лише з деякою упевненістю, або ймовірністю, яку має ця подія.

Імовірність. З точки зору біологічної статистики ймовірність розглядається як числова міра об'єктивної можливості появи випадкової події. У «класичному» розумінні ймовірністю P події A називається відношення числа сприятливих настання цієї події результатів m до числа всіх єдино можливих, рівноможливих і несумісних результатів n випробування:

$$P(A) = m / n \ , \qquad (5.1)$$

де $P(A)$ – ймовірність події A, m – число, що є сприятливим для настання цієї події результатів, n – число всіх єдино можливих, рівноможливих і несумісних результатів випробування.

Чим більше шансів, що сприяють настанню очікуваної події, тим вище його ймовірність.

Імовірність – число, укладене між нулем і одиницею, тобто виражається в частках одиниці, але може бути виражена і у відсотках. При $P=1$ подія називається **достовірною**, тобто, єдино можливим виходом в умовах випробування. При $P=0$ подія називається **неможливою**, тобто таким, яке в умовах випробування завідомо відбутися не може. Якщо ж подія A в даних умовах може відбутися й не відбутися, а при багаторазових випробуваннях воно обов'язково настає, тобто $0<P(A)<1$, то воно називається подією **можливою** чи **випадковою**.

З цих очевидних властивостей ймовірності випливає, що $P(A)+P(\overline{A})=1$, тобто ймовірність події A і ймовірність протилежного події A в сумі дорівнюють одиниці. Для зручності ймовірність очікуваного події прийнято позначати p, тобто тієї ж буквою, якій позначається частість, а ймовірність протилежного події – q, тобто $P(A)=p$ і $P(A)=q$, звідки $p+q=1$.

Імовірність, яку можна вказати до досвіду, називається **апріорною**.

Теоретичне значення частості, тобто $P(m/n)$, навколо якої коливаються емпіричні значення цієї величини, називається **статистичною ймовірністю події** A. Якщо частість події не дуже сильно відхиляється від імовірності, то її можна прийняти як наближеного значення ймовірності очікуваної події.

Закон великих чисел. Можна показати, що частість $p=\left(\frac{m}{n}\right)$ очікуваної події A наближається до його ймовірності у міру збільшення числа випробувань n. У цьому факті виявляється дія статистичного закону великих чисел, теоретичне обґрунтування якому було дано Якобом Бернуллі (1713), а пізніше математиками петербурзької школи на чолі з П. Л. Чебишева. Закон великих чисел стверджує, що частість події A буде як завгодно близькою до його ймовірності, якщо число випробувань необмежено зростає. У більш точної формулюванні теорема Бернуллі, названа законом великих чисел, говорить: ймовірність того, що відхилення частості m/n від імовірності p очікуваної події A в n незалежних випробувань, яка залишається постійною у всій серії випробувань, перевищить будь наперед заданий як завгодно мале число ε, буде прагнути до нуля, якщо число випробувань (n) необмежено зростає:

$$P\left\{\left|\frac{m}{n}-p\right|>\varepsilon\right\}\to 0\ , \qquad (5.2)$$

$$n\to\infty$$

В біологічній статистикою проводилися численні досліди з перевірки цього закону.

5.1 Нормальний розподіл

Випадкові величини. Одним з фундаментальних понять теорії ймовірностей є поняття випадкової величини. Випадкової називається змінна величина, здатна в одних і тих же умовах випробування приймати різні числові значення, що залежать від супутніх випробуванню випадкових причин, які наперед повністю врахувати неможливо. Випадкові величини діляться на дискретні і безперервні. Випадкова величина називається **дискретною**, якщо вона може приймати значення, що виражаються цілими числами. Якщо ж випадкова величина здатна приймати будь-які числові значення, вона називається **неперервною**. Очевидно, поняття дискретної випадкової величини можна поширити на рахункові ознаки, які виявляють дискретне варіювання, тоді як поняття неперервної випадкової величини має поширюватися на мірні ознаки, які варіюють безперервно.

Випадкова величина X в n незалежних повторних випробувань може приймати самі різні значення – $x_1, x_2, x_3, \ldots, x_n$, але в кожному окремому випробуванні вона приймає один з можливих значень. Функція $P(X)$, що зв'язує значення x_i з їх ймовірностями p_i, називається **законом розподілу випадкової величини**, який можна виразити у вигляді таблиці і кривою ймовірності

X	$\ldots x_1$	x_2	$x_3 \ldots x_n$
$P(X)\ldots$	p_1	p_2	$p_3 \ldots p_n$

і описати **формулою Бернуллі**, що дозволяє знаходити ймовірність будь-якого значення цієї величини. Стосовно ж неперервної випадкової величини мова може йти лише про тих значеннях, які вона здатна прийняти з тією чи іншою ймовірністю і інтервалі від – до, який може бути великим і малим.

Закон нормального розподілу. Видатні математики Муавр і Ламберт, Лаплас і Гаус довели, що ймовірність P будь-якого значення (x_i) безперервно розподіляються випадковим величини X знаходиться в інтервалі від x до $x + dx$ (dx – величина, що визначає ширину інтервалу):

$$P(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} dx \ . \qquad (5.3)$$

У цій формулі е = 2,7183 ... – основа натуральних логарифмів; σ – стандартне відхилення, що характеризує ступінь розсіювання значень (x_i) випадкової величини X навколо генеральної середньої μ, яка називається **математичним очікуванням**. У показник ступеня числа e входить нормоване відхилення $t = \left(\frac{x_i - \mu}{\sigma}\right)$. Цей показник, грає важливу роль в дослідженні властивостей нормального розподілу.

Закон нормального розподілу, або просто нормальний закон, описуваний формулою Гауса – Лапласа, висловлює функціональний зв'язок між імовірністю $P(X_i)$ і нормованим відхиленням t. Він затверджує, що ймовірність відхилення будь варіанти (x_i) від центру розподілу μ, де $x_i - \mu = 0$, визначається функцією нормованого відхилення t. Графічно ця функція виражається у вигляді кривої ймовірності, яка називається **нормальною кривою** (рис. 5.1).

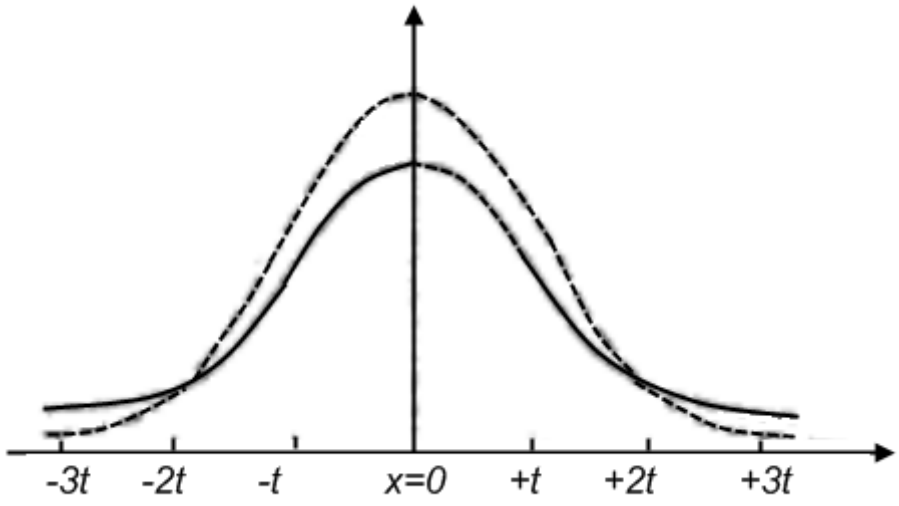


Рисунок 5.1 розподіл прі на тлі кривої нормального розподілу (зображена пунктиром)

Положення цієї кривої повністю визначається двома параметрами: середньою величиною або математичним очікуванням (μ) і стандартним відхиленням (σ), що характеризує варіювання окремих значень випадкової величини навколо центру розподілу μ. У залежності від величини σ форма нормальної кривої може бути пологої (при великій величині σ) у більш-менш крутий (при невеликій величині σ). У всіх випадках нормальна крива строго симетрична щодо центру розподілу і зберігає правильну дзвоноподібну форму (рис. 5.2).

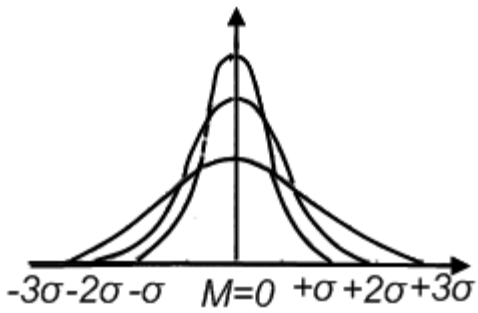


Рисунок 5.2 Криві нормального розподілу з параметрами $M_0 = 0$, $\sigma_1 = 2$, $\sigma_2 = 3$ і $\sigma_3 = 4$.

Якщо стандартне відхилення прирівняти одиниці, тобто $\sigma = 1$, то нормальна крива матиме постійну (стандартизовану) форму, описувану рівнянням

$$y = f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} . \qquad (5.4)$$

Крива, що описується цією формулою, має площу, рівну одиниці (рис. 5.3). Її вершина, тобто максимальна ордината ($y_{\max}$), відповідає початку прямокутних координат, перенесеному у центр розподілу, де $x_i - \mu = 0$.

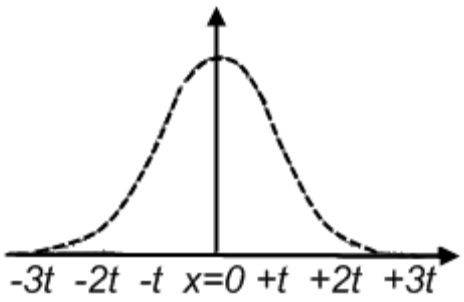


Рисунок 5.3 Стандартизована форма нормальної кривої.

Вправо і вліво від цього центру випадкова величина X може приймати будь-які значення, і ймовірність кожного відхилення $x_i - \mu$, як уже зазначено, визначається функцією його нормованого відхилення $f(t)$. Вірогідність P таких відхилень, що відповідають різним значенням, наводяться в додатку А.

Щоб ордината висловлювала не ймовірності, а абсолютні числові значення випадкової величини, тобто теоретично очікувані частоти варіант емпіричного розподілу, потрібно в праву частину формули (3.17) ввести додаткові множники: у чисельник – загальна кількість спостережень (n), помножене на класовий проміжок (i), а в знаменник – величину середнього квадратичного відхилення вибіркової сукупності (s). У результаті виходить

$$p' = \frac{ni}{s} f(t) \ , \tag{5.5}$$

де p' – теоретично обчислені, або очікувані, частоти варіаційного ряду, $f(t)$ – значення функції нормованого відхилення, розраховані за формулою 5.4.

Користуючись таблицями додатків А та Б, можна по двох емпіричним показниками – середньої арифметичної ($\overline{x}$) і середньому квадратичному відхиленню (s) – обчислити очікувані частоти емпіричного варіаційного ряду, розрахувати ординати і побудувати графік нормальної кривої. Таким чином можна перевірити, чи слід емпіричний ряд розподілу нормальному закону.

Параметри нормального розподілу. Нормальне розподіл повністю характеризується двома параметрами: середньою величиною або математичним очікуванням (μ) і дисперсією випадкової величини X (σ_x^2). Математичне сподівання дискретної випадкової величини дорівнює сумі добутків окремих значень цієї величини на їх ймовірності:

$$\mu(x) = x_1 p_1 + x_2 p_2 + ... + x_n p_n = \sum_{i=1}^{n} x_i p_i \ . \tag{5.6}$$

Формально математичне сподівання відповідає середній арифметичній емпіричних розподілів. Проте ототожнювати ці величини не можна. Середня арифметична виражається відношенням суми всіх членів

ряду до їх загального числа, а математичне очікування становить суму добутків членів ряду на їх вірогідність. Емпірична середня прагне до своєї ймовірної величиною, тобто до математичного сподівання в міру збільшення числа випробувань: чим більше число випробувань, тим ближче емпірична середня до математичного сподівання, і, навпаки, при малих числах випробувань середня арифметична може значно відхилятися від свого математичного очікування.

Дисперсія випадкової величини X дорівнює математичному очікуванню квадрата відхилень окремих значень (x_i) цієї величини від її математичного сподівання (μ), т. е.

$$\sigma_x^2 = \mu\left[x_i - \mu(x)\right]^2 . \qquad (5.7)$$

де σ_x^2 – дисперсія випадкової величини X,

μ – математичне очікування,

x_i – окремі значення величини.

Основні властивості нормального розподілу. Для нормального розподілу характерно збіг за абсолютною величиною середньої арифметичної, медіани і моди. Рівність цих показників вказує на нормальність розподілу випадкової величини. Для нормального розподілу характерно також те, що на рівні інтервали, вимірювані нормованим відхиленням від центру розподілу, доводиться рівне число варіант. Імовірність будь варіанти відхилитися в ту або іншу сторону від середньої μ на t, $2t$ і $3t$, як це видно з додатку А, дорівнює:

$$P\left\{-t < \left|x_i - \mu\right| < +t\right\} = 0,6827;$$

$$P\left\{-2t < \left|x_i - \mu\right| < +2t\right\} = 0,9545;$$

$$P\left\{-3t < \left|x_i - \mu\right| < +3t\right\} = 0,9973.$$

Це означає, що при розподілі сукупності спостережень за нормальним законом з 10000 варіант в інтервалі від $\mu - t$ до $\mu + t$ виявиться 6827 варіант, або 68,3% від обсягу сукупності. В інтервалі від $\mu - 2t$ до $\mu + 2t$ перебуватиме 9545 варіант, або 95,4%, і в інтервалі від $\mu - 3t$ до $\mu + 3t$ виявиться 9973, або 99,7% від числа всіх варіант сукупності. Отже, з

імовірністю $P = 0,6827$ можна стверджувати, що навмання відібрана з генеральної сукупності, що нормально розподіляється варіанту не вийде за границі від $\mu - t$ до $\mu + t$, або в компактній формі $\mu \pm t$. А ймовірність того, що випадково відібрана варіанту не відхилиться від генеральної середньої більш ніж на $\mu \pm 3t$, дорівнює $P = 0,9973$. Це означає, що 99,7% всіх варіант генеральної сукупності, що нормально розподіляється знаходиться в межах $\mu \pm 3\sigma$. Цей важливий висновок відомий в біометрії як правило «плюс – мінус трьох сигм».

5.2 Розподіл рідкісних подій (Закон Пуассона)

Характер біноміальної кривої визначається двома величинами: числом випробувань і ймовірністю очікуваного результату. При $p = 0,5$ біноміальна крива строго симетрична і в міру числа випробуванні набуває більш плавний хід на всьому протязі. Якщо ж $p \neq q$, Біноміальна крива стає асиметричною особливо при збільшенні різниці між p і q. Коли ймовірність очікуваного події обчислюється сотими і тисячними частками одиниці, розподіл частоти такого рідкісного події в n незалежних випробувань виявляється вкрай асиметричним. Розподіл частоти таких рідкісних подій описується формулою Пуассона:

$$P_n(m) = \frac{a^m}{m!} \cdot e^{-n} = \frac{a^m}{m! e^a} \ , \tag{5.8}$$

де $m -$ частота очікуваної події в n незалежних випробувань;

$a \cong np$ – найімовірніше частота рідкісної події;

$e = 2,7183$ ...– основа натуральних логарифмів;

$m!$– факторіал частоти, або добуток натуральних чисел $1 \cdot 2 \cdot 3 ... \ m$.

За формулою Пуассона визначається ймовірність частоти рідкісних подій у серії повторних випробувань.

Щоб формула Пуассона висловлювала не вірогідність, а очікувані абсолютні частоти (p') рідкісної події, їй надається такий вираз:

$$p' = n \frac{\bar{x}^m}{m!} \cdot e^{-\bar{x}} \ . \tag{5.9}$$

Тут p' – теоретичні ординати кривої розподілу Пуассона або очікуване число випадків рідкого події в кожному окремо взятому класі випробування – 0, 1, 2, 3, 4 і т. д.; n – число випробувань: $\overline{x}$ – середнє число фактично спостережуваних випадків (взятого замість n); пояснення інших символів ті ж, що у формулі (5.8).

Розподіл Пуассона – граничний випадок біноміального розподілу. Воно, як і біноміальний розподіл, наближається до кривої ймовірності при зростанні числа $a \cong np$, що видно на рис. 5.4, який ілюструє графік функції $P_n(m)$, побудований для різних значень a.

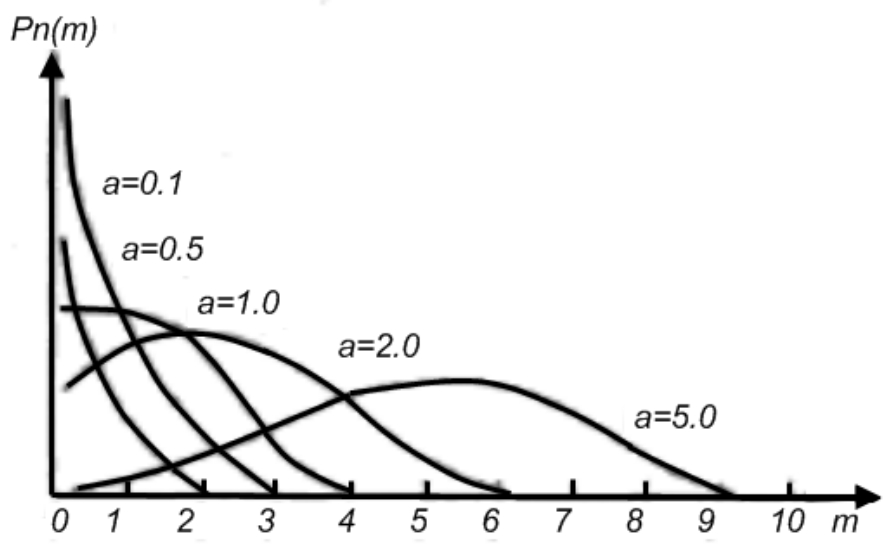


Рисунок – 5.4 Графік функції для різних значень.

За законом Пуассона розподіляються багато випадкові події, з якими доводиться зустрічатися в мікробіології, радіобіології і в інших розділах сучасної біології.

5.3 Біноміальний розподіл

Перш ніж приступити до розгляду одного із законів розподілу ймовірностей, необхідно вивчити правила їх додавання і множення.

1. Імовірність настання одного з двох (все одно якого) або декількох незалежних та несумісних подій $A, B, C, ..., K$ дорівнює сумі їх імовірностей:

$$P(A + B + C + ... + K) = P(A) + P(B) + P(C) + ... + P(K)$$

2. Можливість спільної появи двох або декількох незалежних подій дорівнює добутку ймовірностей цих подій:

$$P(A,B,C,...,K) = P(A)P(B)P(C)...P(K)$$

Наведеними правилами не вичерпуються всі властивості ймовірностей. Більш повні відомості з цього питання вивчає наука теорія ймовірності.

Закон біноміального розподілу. Уявімо, що стосовно події A проводиться n незалежних випробувань. У кожному випробуванні ймовірність появі цієї події постійна – $P(A) = p$. Будемо враховувати тільки два результати: поява A або протилежного йому події $\overline{A}$, теж має постійну ймовірність $P(\overline{A}) = q$, причому $p + q = 1$. За цих умов, якщо подія A в n незалежних випробувань зустрічається m раз, то подія $\overline{A}$ буде зустрічатися $n - m$ раз і ймовірність будь-якого результату, яку позначимо символом $P_n(m)$, незалежно від того, в якому порядку ці події чергуються, виразиться добутком $p^m q^{n-m}$ (за правилом множення ймовірностей), помноженим на біноміальний коефіцієнт $C_n^m = \dfrac{n!}{m!(n-m)!}$, або число сполучень із n елементів по m

$$p_n(m) = C_n^m p^n q^{n-m} \ . \qquad (5.10)$$

Ця формула Бернуллі дозволяє знаходити ймовірність того, що з n узятих навмання елементів виявиться m очікуваних (за умови, що ймовірність очікуваного події дорівнює p).

Сукупність ймовірностей $P_n(m)$ здійснення події A при $0, 1, 2, 3, ..., m$ випадках, тобто $P_n(0), P_n(1), P_n(2), P_n(3), P_n(4), ..., P_n(m)$ називається **біноміальним розподілом**. А так як $p + q = 1$, тоді $\sum_{m=0}^{n} P_n(m) = 1$.

Можна показати, що $\sum_{m=0}^{n} P_n(m) = (p+q)^n$. Так, при $n = 2$ можливі наступні результати: AA $A\overline{A}$ $\overline{A}A$ $\overline{AA}$; всього 2^2=4 результату; їх ймовірності: pp pq qp qq, або $(p+q)^2 = p^2 + 2pq + q^2 = 1$. При трьох

незалежних випробуваннях можливі 2^3=8 фіналів, ймовірності яких розподіляться таким чином: $(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3 = 1$.

При n незалежних випробувань

$$(p+q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{1\cdot 2}p^{n-2}q^2 + ... + q^n .$$

Таким чином, закон біноміального розподілу виражається формулами Бернуллі і Ньютона. За цим законом розподіляються результати очікуваного результату (A) в n незалежних випробувань, коли ймовірність події A дорівнює $p = 0{,}5$. При $n = 10$ можливі 2^{10}=1024 результату.

Очевидно, що розподіл ймовірностей $\sum_{m=0}^{n} P_n(m) = (p+q)^n$ дотримується коефіцієнтів розкладання бінома Ньютона, віднесених до одного й того ж знаменника, рівного 2^n. Біноміальні коефіцієнти легко виходять за допомогою арифметичного трикутника Паскаля, в якому кожна цифра виходить підсумовуванням двох що стоять над нею.

Таким чином, характер біномного розподілу не зміниться від того, як будуть виражені результати випробувань – у значеннях імовірності або в абсолютних значеннях частоти очікуваного результату. У тому і іншому випадку закон розподілу виражає залежність між частотою очікуваного результату і кількістю незалежних випробувань, проведених відносно даної події A, причому частота появи очікуваної події в n незалежних випробувань визначається його ймовірністю, яка залишається постійною в кожному окремому випробуванні.

Закон біномного розподілу піддається експериментальній перевірці.

За біноміальним законом розподіляються багато рахункових і не тільки альтернативні ознаки, тобто ознаки з двома очікуваними наслідками, а й такі, у яких кількість очікуваних наслідків більша за дві.

5.4 Розподіл Максвелла

Поряд з описаними тут типами розподілів випадкових величин у біології зустрічаються не тільки симетричні, але і асиметричні розподілу, які, однак, не підкоряються закону Пуассона. Одним з таких розподілів є розподіл Максвелла:

$$P(x) = \frac{2}{\sqrt{2\pi}} \frac{t^2}{a} e^{-\frac{t}{2}^2} dx \ . \qquad (5.11)$$

У цій формулі a – параметр розподілу, визначається через середнє значення варіюючої ознаки за формулою $a = 0,6267\overline{x}$; $t = x_i / a$ де x_i – числові значення випадкової величини X ; dx – різниця між двома суміжними значеннями змінної величини X .

Вказівкою на те, що емпіричне розподіл слідує закону Максвелла, служить рівність між середнім квадратичним відхиленням і величиною $0,674\,a$, тобто S_x=$0,674\,a$, тоді як розподіл Пуассона характеризується рівністю $S_x^2 = \overline{x}$.

Закони розподілу випадкових величин – це імовірнісні моделі емпіричних розподілів. Вони служать теоретичною основою статистичного аналізу в самому широкому сенсі. Різних типів розподілів багато. Досить поширеним типом розподілу кількісних ознак є нормальний розподіл. Тому нормальний закон особливо важливий у біологічних (і не тільки в біологічних) дослідженнях; він важливий як в теоретичному, так і в прикладному значеннях, зокрема для вироблення нормативів, наприклад, фізичного розвитку людини за тими ознаками, які розподіляються за нормальним законом або не дуже сильно відхиляються від нього (метод сигмальних відхилень). Якщо ж емпіричне розподіл не слід нормальному закону і його не вдається трансформувати (логарифмування значень ознаки) в нормальний ряд, більш точними характеристиками при виробленні нормативів будуть структурні характеристики – медіана, мода і особливо процентильні оцінки.

Питання для самоперевірки:

1. Що називається законом розподілу випадкових величин?
2. Що називається подією?
3. Дайте визначення випадковим подіям.
4. Що таке несумісні події?
5. Що таке сумісні події?
6. Що таке протилежні події?
7. Дайте визначення ймовірності випадкової події.
8. Яка подія називається достовірною?

9. Яка подія називається неможливою?
10. Яка подія називається можливою чи випадковою?
11. Яка ймовірність випадкової події називається апріорною?
12. Що таке статистична ймовірність події?
13. Охарактеризуйте закон великих чисел.
14. Що таке випадкова величина?
15. Яка величина називається дискретною?
16. Яка величина називається неперервною?
17. Охарактеризуйте закон розподілу випадкових величин.
18. Що описує формула Бернуллі?
19. Охарактеризуйте нормальний закон розподілу випадкових величин.
20. Що називається математичним очікуванням випадкової величини?
21. Що таке нормальна крива розподілу випадкової величини?
22. Які форми нормальної кривої Ви знаєте?
23. Назвіть основні параметри нормального розподілу випадкової величини.
24. Перерахуйте основні властивості нормального розподілу випадкових величин.
25. Охарактеризуйте закон Пуассона (розподіл рідкісних подій).
26. Що таке факторіал частоти?
27. Що являють собою теоретичні ординати кривої розподілу Пуассона?
28. Охарактеризуйте закон біноміального розподілу випадкових величин.
29. Перерахуйте правила додавання та множення ймовірностей випадкових величин.
30. Що називається біноміальним розподілом випадкових величин.
31. Охарактеризуйте розподіл Максвелла.

6 ПЕРЕВІРКА ГІПОТЕЗИ ПРО ЗАКОНИ РОЗПОДІЛУ

Перевірка нормальності розподілу за допомогою показників асиметрії та ексцесу. Вибіркові характеристики – середня величина і показники варіації – не містять інформації про закон розподілу генеральної сукупності, з якої вибірка взята. Важко судити про закон розподілу і за емпіричною варіаційної кривої, оскільки на ній позначається вплив численних випадкових причин.

Між тим знання закону розподілу важливо: воно гарантує від можливих помилок в оцінці генеральних параметрів на основі вибіркових показників.

Багато біологічні ознаки розподіляються нормально. Нерідко, однак, емпіричні ряди розподілу відхиляються більш-менш помітно від нормальної кривої. Ці відхилення можуть бути різними, виявляючи в одних випадках *асиметрію*, в інших – *ексцес*, а іноді й те й інше одночасно.

Асиметрія ряду виражається графічно у вигляді скошеної варіаційної кривої, вершина якої може бути зрушена від центру розподілу або вліво, або вправо. Асиметрію називають правобічною чи додатною, якщо вершина кривої перемістити вліво від центру розподілу; вона більш полога, сильно розтягнута по осі абсцис (рис. 6.1). При лівосторонній, або від'ємній, асиметрії, навпаки, вершина кривої зрушена вправо від центру розподілу, а її полога частина знаходиться на лівій стороні (рис. 6.2).

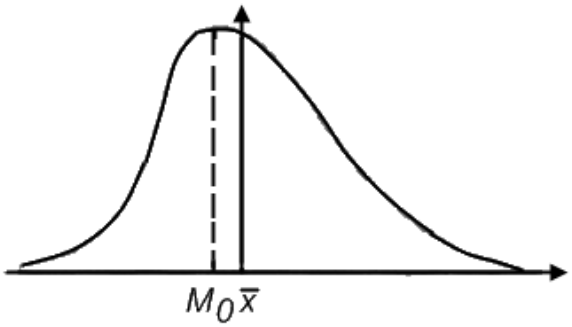


Рисунок 6.1 – Асиметрична крива (додатна асиметрія)

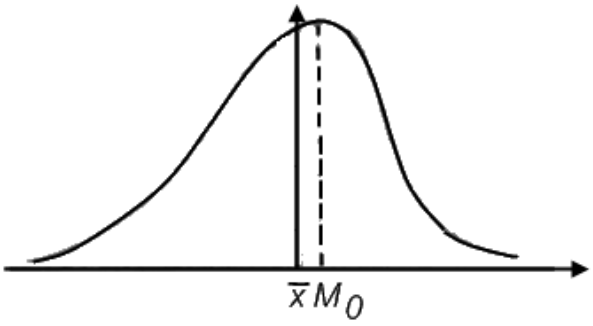


Рисунок 6.2 – Асиметрична крива (від'ємна асиметрія)

Поряд з асиметричними трапляються гостро- і плосковершинні криві розподілу. Гостровершинність спричиняється надмірним накопиченням чисельності варіанта в центрі варіаційного ряду, внаслідок чого вершина кривої різко піднімається (рис. 6.3). Крім одновершинних трапляються дво- і багатовершинні емпіричні криві.

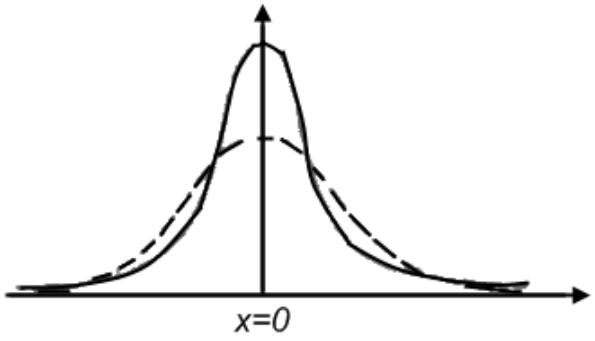


Рисунок 6.3 – Крутовершинний розподіл (додатний ексцес)

Якщо при збільшенні обсягу вибірки плосковершинна крива стає двумодальною, говорять про наявність у такого розподілу від'ємного ексцесу (рис. 6.4).

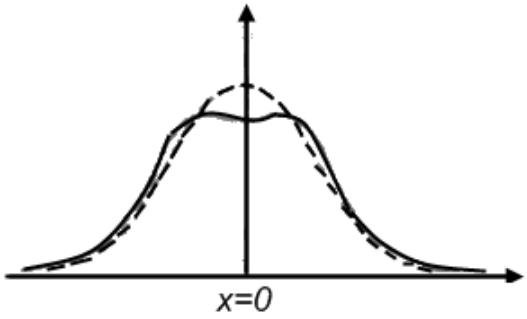


Рисунок 6.4 – Плосковершинний розподіл (від'ємний ексцес)

Асиметрія і ексцес емпіричних розподілів можуть виникнути як наслідок систематично діючих на ознаку (визначених), так і внаслідок випадкових (невизначених) причин. Звідси виникає необхідність у кожному випадку встановлювати, випадкові або не випадкові відхилення емпіричних розподілів від нормальної кривої Наближено оцінювати нормальність розподілу дозволяють центральні моменти третього і четвертого порядків, що використовуються для вимірювання асиметрії та ексцесу. **Показник асиметрії**, що позначається символом As, представляє **центральний момент третього порядку**, віднесений до куба середнього квадратичного відхилення: $As = \mu_3 \big/ \sigma_x^3$, а його оцінкою служить наступна величина:

$$As = \frac{\sum_{i=1}^{k} p_i (x_i - \overline{x})^3}{nS_x^3} \ . \qquad (6.1)$$

При строго симетричних розподілах сума третього ступенів відхилень варіант (x_i) від середньої арифметичної ($\overline{x}$) дорівнює нулю. При наявності ж скошеності розподілу, цей показник буде мати або позитивну, або негативну величину, яка і служить мірою асиметрії. При правобічної асиметрії будуть переважати куби відхилень з позитивним знаком, а при лівобічної асиметрії – з від'ємним. Звідси і коефіцієнт асиметрії буде мати додатний або негативний знак. При відсутності асиметрії $As = 0$.

Показник ексцесу, що позначається символом Ex, виражається формулою $Ex = \mu_4 \big/ \sigma_x^4 - 3$, а його оцінкою служить наступна величина.

$$Ex = \frac{\sum_{i=1}^{k} p_i (x_i - \overline{x})^4}{nS_x^4} - 3 \ . \qquad (6.2)$$

При відсутності ексцесу $Ex = 0$. Якщо ексцес додатний, то цей показник набуває додатний знак будь-яку величину, так як теоретично нічим не обмежений. При плосковершинні коефіцієнт Ex має негативний знак; гранична величина від'ємного ексцесу дорівнює – 2.

Як і інші оцінки генеральних параметрів, показники асиметрії та ексцесу є величинами випадковими і супроводжуються помилками репрезентативності, які визначаються за наступними наближеним формулами:

$$S_{As} = \sqrt{\frac{6}{n+3}} \ , \qquad (6.3)$$

$$S_{Ex} = \sqrt{\frac{24}{n+5}} = 2\sqrt{\frac{6}{n+5}} \qquad (6.4)$$

Більш точні формули помилок цих показників наступні:

$$S_{As} = \sqrt{\frac{6\,(n-1)}{(n+1)(n+3)}} = \sqrt{\frac{6n\,(n-1)}{(n-2)(n+1)(n+3)}} \, ;$$

$$S_{Ex} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}} \, .$$

Нульова гіпотеза, або припущення, що в генеральній сукупності показники і рівні нулю, спростовується якщо:

$$t_{As} = \frac{As}{S_{As}} > 3$$

або

$$t_{Ex} = \frac{Ex}{S_{Ex}} > 3.$$

Обчислення показників As і Ex основним способом, тобто через центральні моменти третього і четвертого порядків, виявляється досить трудомістким. Тому частіше ці показники розраховують за способом умовної середньої, тобто використовують умовні моменти, які певним чином пов'язані з центральними моментами.

Звідси випливають такі робочі формули:

$$As = \left(\frac{\sum pa^3}{n} - 3b_1 \frac{\sum pa^3}{n} + 2b_1^3 \right) \Big/ S^3 \ , \qquad (6.5)$$

$$Ex = \left[\left(\frac{\sum pa^4}{n} - 4b_1 \frac{\sum pa^4}{n} + 6b_1^2 \frac{\sum pa^4}{n} - 3b_1^4 \right) \Big/ S^4 \right] - 3 \ . \qquad (6.6)$$

де $a-$ відхилення варіант (випадкових величин x_i) від умовної середньої (A), тобто $a = x_i - A$; $b_1 = \sum p(x_i - A)/n$.

6.1 Розрахунок теоретичних частот

За нормальним законом. Для перевірки гіпотези про закон розподілу, якому слід емпірична сукупність, необхідно частоти емпіричного розподілу зіставити з теоретично обчисленими частотами. Останні розраховуються на основі емпіричних даних за формулами, які описують той чи інший закон розподілу ймовірностей. Так, при перевірці нормальності розподілу теоретичні частоти розраховуються за формулою (5.5), згідно з якою значення функції $f(t)$, наведені для різних значень

нормованого відхилення (t) в додатку Б, потрібно помножити на величину ni/S_x, де n – обсяг вибірки; i – величина класового інтервалу або проміжок між сусідніми класами емпіричного варіаційного ряду, S_x – середнє квадратичне відхилення цього ряду.

Якщо емпіричні і обчислені за нормальним законом частоти цього розподілу зобразити графічно у вигляді варіаційних кривих, можна наочно переконатися у хорошому збігу їх один з одним.

Визначення теоретичних частот емпіричного ряду з того чи іншого закону розподілу ймовірностей є процес вирівнювання емпіричного розподілу, звільнення його від усього випадкового, з тим щоб виявити основну тенденцію варіювання ознаки, закон його розподілу.

За біноміальним законом. Щоб перевірити, чи відповідає емпіричне розподіл дискретно варіюючої ознаки біноміальним законом, потрібно, як і в попередньому випадку, обчислити теоретичні частоти цього розподілу і порівняти їх з частотами емпіричного ряду. Для розрахунку теоретичних частот за біноміальним законом служить наступна формула:

$$p' = N(p+q)^{n-1}\ , \tag{6.7}$$

яка в кінцевому рахунку приймає вигляд

$$p' = 100(qpK)\ , \tag{6.7 а}$$

тут p – емпірична ймовірність, або частка середнього результату, яка визначається за формулою

$$p = \frac{\bar{m}}{n-1} \tag{6.8}$$

де $\bar{m} = \frac{\sum mp_i}{N}$ – середня арифметична, а $N = \sum p_i$ – сума частот емпіричного ряду, або обсяг вибірки; $q = 1 - p$, а K – відповідний коефіцієнт біноміального ряду $(1+1)^{n-1}$.

Ця формула особливо зручна для розрахунку теоретичних частот на моделях з невідомою ймовірністю.

Коли відома або може бути встановлена апріорі ймовірність p досліджуваного явища, тобто за наявності біноміального розподілу з відомою ймовірністю, теоретичні частоти можна розрахувати за формулою:

$$p' = \frac{KN}{\sum K}\ , \qquad (6.9)$$

де K і N теж значення, що й у формулах (6.7) і (6.7 а).

За законом Пуассона. Розрахунок теоретичних частот (p') здійснюється за значенням ймовірності $P_n(m) = \frac{a^m}{m!} \cdot e^{-a}$ тобто величині, що входить до складу формули Пуассона і визначальною функцію розподілу частот рідкісних подій. Ці значення, відповідні частотах рідкісних подій, містяться в додатку В. Техніка розрахунку теоретичних частот за законом Пуассона дуже проста. Визначивши середню арифметичну $(\overline{x})$ емпіричного розподілу, в додатку В знаходять ймовірності частот m: 0, 1, 2, 3 і т. д. до кінця ряду. Потім ймовірності частот множать на загальне число спостережень (n), що і дає в результаті значення теоретичних частот (p') для кожного значення m.

У результаті вийшов ряд теоретичних частот, який непогано узгоджується з частотами емпіричного ряду

За законом Максвелла. Для розрахунку теоретичних частот за формулою Максвелла (див. формулу (5.11)) надходять у такий спосіб.

1. Знаходять середню арифметичну емпіричного варіаційного ряду $(\overline{x})$ і параметр $a = 0.6267 \cdot \overline{x}$.
2. Поділом кожної класової варіанти (x_i) на величину a, визначають значення $t = x_i / a$.
3. За додатком Б для кожного значення $t = x_i / a$ знаходять $f(t)$.
4. Визначають відношення t^2 / a.
5. Знайдені значення t^2 / a множать на подвоєну величину функції t, тобто на $2f(t)$, і величину класового проміжку $(i = dx)$.
6. Отримані величини $t^2 2f(t)i / a = P$ множать на загальне число спостережень (n), що і дає теоретичні частоти ряду $(Pn = p')$.

У результаті вийшов ряд теоретично обчислених частот (p'), який непогано узгоджується з емпіричними частотами цього розподілу.

6.2 Критерій відповідності емпіричних частот частотам обчисленим або очікуваним

Критерій χ^2 («хі-квадрат») К. Пірсона. Як би точно ні обчислювалися теоретичні частоти, вони як правило, не збігаються з емпіричними частотами ряду. Звідси виникає необхідність порівняння емпіричних частот з обчисленими, або очікуваними, частотами, з тим щоб встановити достовірність або випадковість спостережуваного між ними розбіжності. Нульова гіпотеза зводиться до припущення, що невідповідність емпіричних частот частотам, обчисленими з того чи іншого закону розподілу, – цілком випадкове, тобто між обчисленими і емпіричними частотами ніякої різниці немає. Для перевірки нульової гіпотези використовуються особливі критерії. Одним з найбільш часто вживаних в біологічній статистикою служить критерій χ^2, запропонований К. Пірсоном у 1900 р. Цей критерій являє суму квадратів відхилень емпіричних частот (p) від частот теоретичних або очікуваних (p'), віднесену до теоретичних частотах (p'), тобто

$$\chi^2 = \sum_{i=1}^{k} \frac{(p - p')^2}{p'} . \qquad (6.10)$$

Символ χ^2 – не квадрат якогось числа, він висловлює лише вихідну величину, яка визначається даною формулою. Позначивши різниця між емпіричними і теоретичними частотами через d, можна написати:

$$\chi^2 = \sum_{i=1}^{k} \frac{d}{p'} . \qquad (6.11)$$

Так як відхилення емпіричних частот від очікуваних чи обчислених зводяться в квадрат, величина критерію χ^2 завжди додатна. Тому при визначенні різниці $(p - p') = d$ знаки можна не враховувати, віднімаючи з великих чисел менші.

При повному збігу емпіричних частот з частотами розрахунковими чи очікуваними $\sum(p - p') = 0$ і критерій χ^2 теж буде дорівнює нулю. Якщо ж $\sum(p - p') \neq 0$, це вкаже на невідповідність обчислених частот емпіричним частотах ряду.

У таких випадках необхідно оцінити значущість критерію χ^2, який теоретично може змінитися від 0 до ∞.

Це проводиться шляхом порівняння фактично отриманої величини $\chi^2_{\hat{o}}$ з його критичним значенням (χ^2_{st}). Нульова гіпотеза, тобто

припущення, що розбіжність між емпіричними і теоретичними чи очікуваними частотами носить випадковий характер, спростовується, якщо $\chi^2_{\hat{o}} \geq \chi^2_{st}$ для прийнятого рівня значущості (α) та кількості степенів вільності (k). Критичні значення χ^2 для різних рівнів значущості і чисел степенів вільності містяться в додатку Г.

Розподіл ймовірних значень випадкової величини безперервно і асиметрично (рис. 6.5). Воно залежить від числа степенів вільності(k) і наближається до нормального розподілу в міру збільшення числа спостережень (n).

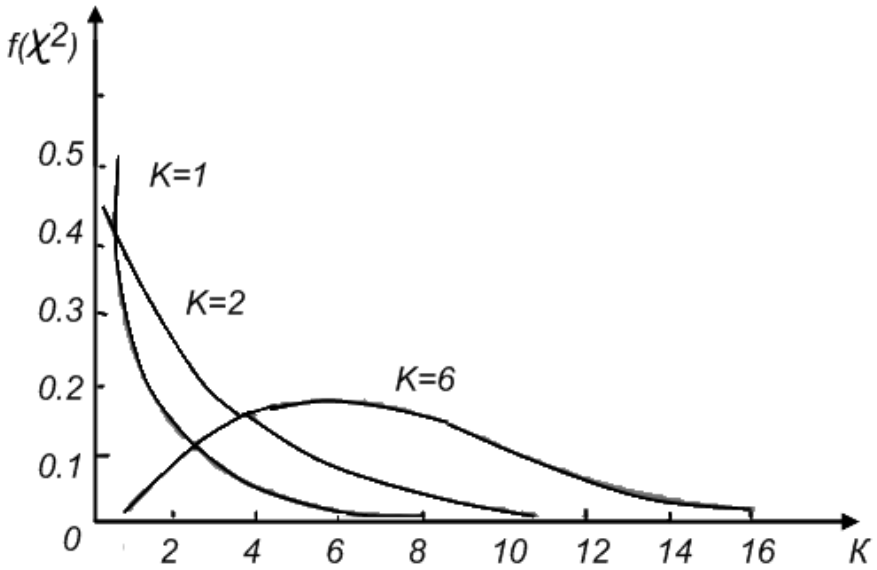


Рисунок 6.5 – Функція хі-квадрат розподілу в залежності від різних чисел степеня вільності (k).

Тому застосування критерію χ^2 до оцінки дискретних розподілів пов'язане з деякими похибками, які позначаються на його величині, особливо на нечисленних вибірках. Для отримання більш точних оцінок вибірка, що розподіляється в варіаційний ряд, повинна мати не менше 50 варіант. Правильне застосування критерію вимагає також, щоб частоти варіант у крайніх класах не були б менше 5; якщо їх менше 5, то вони об'єднуються з частотами сусідніх класів, щоб в сумі складали величину, більшу, або рівну 5. Відповідно об'єднанню частот зменшується і число класів (N). Число ступенів волі встановлюється по вторинному числа класів з урахуванням числа обмежень свободи варіації. Так, при оцінці емпіричного розподілу по нормальному або біноміальним законом існує при обмеження свободи варіації: $\overline{x}$, S_x і n, а число степенів

вільності $k = N - 3$. Якщо ж оцінка здійснюється за законом Пуассона, для якого відомі два обмеження: n і $\bar{x}$ (або S_x^2), число степенів вільності $k = N - 2$. В інших випадках число ступенів волі (k) встановлюється особливо.

Так як точність визначення критерію χ^2 в значній мірі залежить від точності розрахунку теоретичних частот (p'), для отримання різниці між емпіричними і обчисленими частотами $(p - p') = d$ слід використовувати неокруглені теоретичні частоти (p').

Критерій χ^2 можна використовувати і для порівняння емпіричних рядів з їх частотами, які розподілені по одним і тим же класах. У таких випадках застосовується наступна формула:

$$\chi^2 = \frac{1}{n_1 n_2} \sum \frac{(n_1 p_2 - n_2 p_1)^2}{p_1 + p_2} \ , \qquad (6.12)$$

де n_1 і n_2 – обсяги порівнюваних вибірок, розподілені в варіаційні ряди, p_1 і p_2 – частоти першого і другого рядів розподілу. Нульова гіпотеза зводиться до того, що порівнювані вибірки взяті з однієї і тієї ж генеральної сукупності і, отже, розбіжність між частотами p_1 і p_2, носить випадковий характер. Як і в попередніх випадках, при наявності в крайніх класах менше 5 варіант їх необхідно об'єднати з частотами сусідніх класів і по вторинному числа визначати число степенів вільності (k).

Критерій χ^2 застосовується і тоді, коли емпіричні та очікувані частоти групуються у чотирипільну чи багатопільні таблиці. У таких випадках число степенів вільності визначається за кількістю рядків і стовпців, без урахування підсумків таблиці, за наступною формулою:

$$k = (c - 1)(\tilde{a} - 1) \ , \qquad (6.13)$$

де $c -$ число сторінок, $\tilde{a} -$ кількість граф, або стовпців, таблиці.

При цьому критерій χ^2 непридатний, якщо результати спостережень виражені не абсолютними, а відносними числами у відсотках, частостях і інше.

Приклад таблиці, яка використовується в даному випадку:

Классы	Частоты	$p_1 + p_2$	$n_1 p_2$	$n_2 p_1$	$n_1 p_2 - n_2 p_1$	$(n_1 p_2 - n_2 p_1)^2 / (p_1 + p_2)$

(x_i)	p_1	p_2					
Сумма							

Якщо вибіркові дані (наприклад, дані досвіду та контролю) групуються в чотирьохпільну таблицю, критерій χ^2 визначається за формулою

$$\chi^2 = \frac{n \cdot \left(|ad - bc| - \frac{n}{2} \right)^2}{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)} , \qquad (6.14)$$

де a,b,c,d – чисельності груп, поміщені в клітинах чотирьохпільної таблиці, $n = a + b + c + d$ – загальне число спостережень. Прямі дужки, в яких укладена різниця $ad - cb$, вказують на те, що беруться абсолютні різниці, з яких віднімається величина $n/2$ – поправка на безперервність варіації. Якщо не внести цю поправку, то величина критерію χ^2, який має не перериваним функцію розподілу, виявиться трохи завищеною при обчисленні на дискретних і особливо нечисленних вибірках.

Використання формули (6.14) ускладнює обчислювальну роботу. Це незручність можна подолати, якщо скористатися формулою (6.10), внісши до неї поправку на безперервність варіації:

$$\chi^2 = \sum_{i=1}^{k} \frac{(d - 0.5)^2}{p'} , \qquad (6.15)$$

де $d = (p - p')$ – різниця між (p), що спостерігаються і очікуваними або обчисленими (p') чисельностями груп, які розраховуються за формулою:

$$p' = \frac{n_c n_{\tilde{a}}}{n} , \qquad (6.16)$$

де n_c – підсумкові числа по рядку; $n_{\tilde{a}}$ – підсумкові числа по стовпцю чотирьохпільної або багатопільної таблиці, n – загальне число спостережень.

Критерій Ястремського. Піддавши критичному аналізу критерій χ^2 Пірсона, Б.С. Ястремський (1940) прийшов до висновку, що цей критерій не дає основи для судження про ступінь близькості емпіричних частот частотах обчисленими. Він запропонував свій критерій $\pm 3\sqrt{2N+40}$, який виражається у вигляді такої формули:

$$I = \frac{|C - N|}{\sqrt{2N + 40}}\ , \qquad (6.17)$$

Тут I-критерій Ястремського; $C = \sum \frac{(p - p')^2}{p'q}$ де p – фактичні, а p' – обчислені частоти; N – число груп або класів варіаційного ряду, причому $|C - N|$ береться без урахування знака; θ – величина, яка задовольняє нерівності $0.5 \le \theta \le 1$, залежить від числа груп (N) і при $N \le 20$ не перевершує 0,6. Оскільки число класів зазвичай не перевищує 20, величину 4 θ можна вважати рівною 2,4; I розподіляється нормально. Тому з імовірністю $P = 0.997$ можна стверджувати, що критерій I не перевищить ± 3. Це означає, що при $I \le 3$ є підстава стверджувати про нормальний розподіл емпіричної сукупності. При цьому умови нульову гіпотезу відкинути не можна.

6.3 Причини асиметрії емпіричних розподілів

Кетле і Гальтон вважали, що біологічні ознаки, розподіляються нормально, і численні факти начебто б підтверджували це. Але незабаром Пірсон показав, що існує кілька типів розподілів. Виникла необхідність з'ясувати причини асиметрії в розподілі біологічних ознак. Вирішення цього питання має свою історію. В даний час можна вказати на наступні три причини виникнення асиметрії в розподілі біологічних ознак.

Перша полягає в «неправильному» групування вибіркових даних.

Дуже важливо дотримуватися правил групування вибіркових даних в варіаційні ряди, а також стає зрозумілим і необхідність всебічного аналізу кожного емпіричного розподілу з точки зору того чи іншого закону розподілу ймовірностей.

Друга причина асиметричних розподілів – умови зовнішнього середовища, в яких формуються індивіди та їх ознаки.

Третя причина асиметричних розподілів – генетична, вона обумовлена взаємодією алельних і неалельних активних генів. Відомо, що кількісні ознаки успадковуються полігенно; дію на ознаку адитивних (однозначних або схожих за силою дії) генів при відсутності домінування, епістазу та інших способів взаємодії неалельних генів обумовлює проміжний тип успадкування ознаки, його нормальний розподіл. Якщо ж у процесі формування ознаки відбувається взаємодія генів, при якому одні активні алелі здатні пригнічувати або активізувати інші алелі, то не проміжного успадкування, ні повного домінування у спадкуванні ознаки спостерігатися не буде і крива розподілу такого ознаки виявиться асиметричною.

Зрозуміло, що кожна з названих причин діє не ізольовано, а швидше за все спільно з іншими причинами, що викликають асиметрію емпіричних розподілів. Тому без статистичного, а головним чином генетичного, аналізу не можна вказати на конкретну причину або причини відхилення емпіричних розподілів від нормальної кривої. У всякому разі відхилення від нормального закону, якщо вони не випадкові, можуть вказувати на постійно діючу на ознаку причину або причини, з'ясування яких – більше завдання біології, а не біологічної статистики.

Вимірювання трансгресії. При розподілі незалежних вибірок, взятих з різних генеральних сукупностей, нерідко трапляється, що якась частина членів цих вибірок виявляється в одних і тих же класах варіаційного ряду.

Ряди, у яких частина класів виявляються загальними незважаючи на те, що між середніми арифметичними цих рядів різниця може бути статистично достовірної, називаються **трансгресуючими**, а факт неповного роз'єднання варіаційних рядів та їх графіків – **трансгресією**. Варіаційні криві трансгресуючих рядів виглядають так, що права сторона одній кривій і ліва сторона інший взаємно проникають один в одного і під ними утворюється частина загальної площі, що показує величину трансгресії.

Трансгресія рядів може бути різною, що має певне пізнавальне значення.

Величину трансгресії можна виміряти, виразивши суму варіантів, що трансгресують у відсотках від загальної чисельності обох вибірок.

Наведений спосіб вимірювання величини трансгресії не точний і дає лише наближене уявлення про її розмір. Більш точно величина трансгресії для нормального розподілу вимірюється за допомогою наступної формули:

$$T = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \ , \qquad (6.18)$$

де n_1 і n_2 – обсяги зіставляються розподілів; $p_1 = 0.5 + 0.5\hat{o}(t_1)$ і $p_2 = 0.5 + 0.5\hat{o}(t_2)$.

Тут

$$t_1 = \frac{\min_2 - \bar{x}_1}{S_1} \quad \text{і} \quad t_2 = \frac{\min_1 - \bar{x}_2}{S_2}$$

а

$$\min_1 = \bar{x}_1 + 3S_1 \quad \text{і} \quad \min_2 = \bar{x}_2 - 3S_2$$

де $\bar{x}_1$ і $\bar{x}_2$ – середні арифметичні, S_1 і S_2 – середньо квадратичні відхилення.

Показник трансгресії (T) виражається в частках одиниці або відсотках. Значення функції $\hat{o}(t)$ знаходиться в додатку Д, причому їх потрібно брати з від'ємним знаком, якщо $\min_1 < \bar{x}_2$ і $\min_2 > \bar{x}_1$, тобто виходити з $p = 0.5 - 0.5\hat{o}(t)$.

Питання для самоперевірки:

1. За допомогою яких показників проводиться перевірка нормальності розподілу випадкових величин?
2. Що таке асиметрія?
3. Що таке ексцес?
4. Як графічно виражається додатна асиметрія?
5. Як графічно виражається від'ємний асиметрія?
6. Як графічно виражається додатний ексцес?
7. Як графічно виражається від'ємний ексцес?
8. Що є показником ексцесу?
9. Що таке нульова гіпотеза?
10. Як пов'язані асиметрія і ексцес з нульовою гіпотезою.
11. Як ведеться розрахунок теоретичних частот за нормальним законом?
12. Як ведеться розрахунок теоретичних частот за біноміальним законом розподілу?
13. Як ведеться розрахунок теоретичних частот за законом Пуассона?

14. Як ведеться розрахунок теоретичних частот за законом Максвелла?

15. Опишіть критерій χ^2. Для чого він потрібен?

16. Як графічно представляється функція хи-квадрат в залежності від різних чисел степенів вільності?

17. Від чого залежить точність визначення критерію χ^2?

18. За яких умов застосовується критерій χ^2?

19. Опишіть критерій Ястремського. Для чого він потрібен?

20. Чи мають зв'язок критерій χ^2 та критерій Ястремского?

21. Назвіть основні причини асиметрії емпіричних розподілів.

22. Що називається групуванням вибіркових даних?

23. Що називається трансгресією?

24. Що називається трансгресуючими рядами?

25. Назвіть основний спосіб вимірювання величини трансгресії.

7 КОРЕЛЯЦІЙНИЙ АНАЛІЗ

У природі діє єдиний закон загального зв'язку, і залежність, яка спостерігається між біологічними ознаками, – це лише окремі випадки прояву цього закону. Тому природно прагнення використовувати цей закон в інтересах людини, вивчити умови, при яких проявляється його дію, надати йому точний кількісний вираз. Цій меті служить математичне поняття функції, що має на увазі випадки, коли певному значенню однієї (незалежної) змінної X, яка називається **аргументом**, відповідає певне значення іншої (залежної) змінної Y, що називається **функцією**. Однозначна залежність між змінними величинами і називається функцією. Однозначна залежність між змінними величинами Y і X і називається **функціональною**, тобто $Y = f(X)$.

Прикладів функціональних зв'язків багато. Відомо, наприклад, що між довжиною тіла й масою риби існує додатний зв'язок: більш довгі індивіди мають зазвичай і велику масу, ніж індивіди меншою довжини. Проте з цього правила є винятки, коли порівняно низькою довгі індивіди виявляються важче довгих.

Причина таких «виключень» в тому, що кожен біологічний ознака, висловлюючись математичною мовою, є функцією багатьох змінних; на його величині позначається вплив і генетичних і середовищних факторів, в тому числі і випадкових, що викликає варіювання ознак. Звідси залежність між ними набуває не функціональний, а статистичний характер, коли певному значенню однієї ознаки, що розглядається в якості незалежної змінної, відповідає не одне і те ж числове значення, а ціла гама розподіляються в варіаційний ряд числових значень іншої ознаки, що розглядається в якості незалежної змінної . Такого роду залежність між змінними величинами називається **кореляційною** або **кореляцією**.

Термін «кореляція» (від лат. *Correlatio* – співвідношення, зв'язок) вперше застосував Ж. Кюв у праці «Лекції з порівняльної анатомії» (1800). Математичне обґрунтування методу було дано в 1846 р. іншим французьким ученим Огюстом Браве. Однак Браве, обґрунтовуючи метод, мав на увазі «теорію помилок в площині», він переносив закон помилок Гауса на випадки двох змінних Y і X в область кристалографії, якою займався. Розвиток теорії кореляції і застосування її до вивчення спадковості і мінливості кількісних ознак пов'язано з іменами Ф. Гальтона, К. Пірсона та ін Термін "кореляція" ввів в біометрію або біологічну статистику Гальтон (1886).

Якщо функціональні зв'язки однаково легко виявити на одиничних і на групових об'єктах, то цього не можна сказати про зв'язки кореляційних,

які вивчаються тільки на групових об'єктах методами математичної статистики.

Залежність між змінними величинами Y і X можна виразити аналітично за допомогою формул і рівнянь і графічно у вигляді геометричного місця точок у системі прямокутних координат. Графік кореляційної залежності будується за рівнянням функції $\overline{y}_x = f(x)$ і $\overline{x}_y = f(x)$, які називаються регресією. Тут $\overline{y}_x$ і $\overline{x}_y$ – середні арифметичні з числових значень залежних змінних Y і X.

Кореляційний зв'язок між ознаками може бути ***лінійною*** та ***криволінійною (нелінійною)***, ***додатною*** і ***від'ємною***. **Завдання кореляційного аналізу** зводиться до встановлення напрямки і форми зв'язку між ознаками, вимірюванню її тісноти й оцінці достовірності вибіркових показників кореляції.

7.1 Коефіцієнт кореляції

Спряженість між змінними величинами Y і X можна встановити, зіставляючи числові значення однієї величини з відповідними значеннями іншого. Якщо при збільшенні однієї змінної збільшується інша, в наявності додатний зв'язок, і, навпаки, коли збільшення однієї змінної супроводжується зменшенням значень іншої, перед нами негативний зв'язок. Питання це вирішується просто при наявності однозначних зв'язків між змінними величинами, коли мова йде про збільшенні або зменшенні функції за заданим значенням аргументу.

Інша справа варіюють ознаки. Тут доводиться зустрічатися не з приростом або зменшенням функції, а з сполученої варіацією, висловлюючи її у вигляді взаємно пов'язаних відхилень варіант від їх середніх. Віднесенням суми добутків відхилень одного ряду (X) на відповідні відхилення іншого (Y) отримують показник, званий коваріацією (Cov):

$$Cov = \frac{1}{n}\left[\sum (x_i - \overline{x})(y_i - \overline{y})\right].$$

Цей показник характеризує ступінь спряженості між двома варіюючими ознаками Y і X. Недолік його полягає в тому, що він не враховує випадки, коли корелюється ознаки виражаються різними одиницями. Наприклад, маса організму може корелювати з його лінійними розмірами, довжина тіла риби – з масою містяться в них жиру і т. д. Недолік, властивий показником коваріації, усувається, якщо замість відхилень $(x_i - \overline{x})$ і $(y_i - \overline{y})$ взяти їх відносини до середніх квадратичних

відхилень S_x і S_y. У результаті виходить показник, який позначається буквою r і називається **емпіричним коефіцієнтом кореляції**:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x \cdot S_y} \quad . \tag{7.1}$$

Цю формулу можна видозмінити і представити в наступному вигляді:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad . \tag{7.2}$$

Цінність цієї пірсоновскої формули полягає в тому, що вона дозволяє визначати коефіцієнт кореляції, не вдаючись до розрахунку середніх квадратичних відхилень, що спрощує обчислювальну роботу.

Коефіцієнт кореляції – зручний показник зв'язку, що набув широкого застосування в практиці. Це абстрактне число, що лежить в межах від -1 до $+1$. При незалежному варіюванні ознак, коли зв'язок між ними відсутній, $r = 0$. Чим сильніше зв'язок між ознаками, тим більше і величина коефіцієнта кореляції. Отже, при $r > 0$ цей показник характеризує не тільки наявність спряженості між ознаками, але й ступінь її. При позитивній, або прямий, зв'язку, коли великим значенням однієї ознаки відповідають більші значення іншого, коефіцієнт кореляції набуває позитивний (+) знак і знаходиться в межах від 0 до +1, а при негативній, чи зворотної, зв'язку, коли великим значенням однієї ознаки відповідають менші значення іншого, коефіцієнт кореляції супроводжується від'ємним (–) знаком і знаходиться в межах від 0 до -1.

Лише один недолік є у цього цінного показника: він здатний характеризувати тільки лінійні зв'язки, тобто такі, які виражаються рівнянням лінійної функції. При наявності нелінійної залежності між варіюючими ознаками слід використовувати інші показники зв'язку.

7.2 Обчислення коефіцієнта кореляції

Обчислення коефіцієнта кореляції на нечисленних вибірках проводиться безпосередньо за значеннями сполучених ознак, тобто без групування вибіркових даних в варіаційні ряди. Для цього служать наведені основні формули (7.1) і (7.2). Зручніше, проте, особливо в тих

випадках, коли відхилення від середніх виражаються багатозначними до дробовими числами, обчислювати коефіцієнт кореляції за однією з наступних робочих формул:

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}\ , \qquad (7.3)$$

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{D_x D_y}}\ , \qquad (7.4)$$

$$r_{xy} = \frac{D_x + D_y - D_d}{2\sqrt{D_x D_y}}\ , \qquad (7.5)$$

де

$$D_x = \sum (x_i - \overline{x})^2 = \sum x_i^2 - (\sum x_i)^2 / n\,,$$

$$D_y = \sum (y_i - \overline{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n\,,$$

$$D_d = \sum d_i^2 - (\sum d_i)^2 / n\,.$$

Тут x_i і y_i – парні значення сполучених ознак Y і X; $\overline{x}$ і $\overline{y}$ – середні арифметичні; $d = (x_i - y_i)$ – різниця між парними (сполученими) варіантами; n – загальна кількість парних спостережень, або обсяг вибірки.

Використання цих формул вимагає попередньо розрахувати такі допоміжні величини: $\sum x_i$, $\sum y_i$, $\sum x_i y_i$, $\sum x_i^2$, $\sum y_i^2$, а також $\sum d_i$ і $\sum d_i^2$, залежно від того, яка формула застосовується для обчислення коефіцієнта кореляції.

7.3 Оцінка достовірності коефіцієнта кореляції

Емпіричний коефіцієнт кореляції служить оцінкою генерального параметра ρ і як випадкова величина супроводжується помилкою. Остання визначається за формулою

$$S_r = \frac{1-r^2}{\sqrt{n}}, \tag{7.6}$$

або коли обсяг вибірки не перевищує 100 спостережень,

$$S_r = \frac{\sqrt{1-r^2}}{\sqrt{n-2}} = \sqrt{\frac{1-r^2}{n-2}}. \tag{7.7}$$

Ставлення вибіркового коефіцієнта кореляції до своєї помилку служить критерієм для перевірки нульової гіпотези – пропозиції про те, що в генеральній сукупності цей показник дорівнює нулю, тобто $\rho = 0$. Нульова гіпотеза спростовується, якщо

$$t_{\hat{o}} = \frac{r\sqrt{n}}{1-r^2} \geq t_{St}$$

або

$$t_{\hat{o}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = r\sqrt{\frac{n-2}{1-r^2}}$$

(при $n \leq 100$) для $k = n-2$ і прийнятого рівня значущості (α).

Нульова гіпотеза спростовується, якщо емпіричний коефіцієнт кореляції перевищить вказану табличну величину для прийнятого рівня значущості та кількості степенів вільності $k = n-2$.

Встановлено, що при малому обсязі вибірки емпіричний коефіцієнт кореляції (r) виявляється дещо нижче, ніж генеральний параметр (ρ). Тому найкраща оцінка ρ виходить за формулою:

$$r^* = r\left[1 + \frac{1-r^2}{2(n-3)}\right]. \tag{7.8}$$

Зрозуміло, що при наявності великої кількості спостережень ($n \leq 100$) ця поправка виявляється незначною і її можна не вносити.

7.4 z – перетворення Фішера

Правильне застосування коефіцієнта кореляції передбачає нормальність розподілу двовимірної сукупності сполучених значень випадкових змінних велич Y і X. З математичної статистики відомо, що при малому числі випробувань в порівняно сильної кореляції ($r > 0,5$) розподіл коефіцієнта кореляції $n-$го числа вибірок, взятих з сукупності, що нормально розподіляється, значно відхиляється від нормальної кривої.

Маючи на увазі що емпіричний коефіцієнт кореляції не буде точною оцінкою генерального параметра (ρ), якщо він обчислений на нечисленною вибірці і його величина значно відхиляється від 0.5 Фішер знайшов більш точний спосіб оцінки генерального параметра ρ за величиною вибіркового коефіцієнта кореляції r. Цей спосіб зводиться до заміни коефіцієнта кореляції перетвореної величиною, яка пов'язана з емпіричним коефіцієнтом кореляції наступним чином:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

або

$$z = 1,15129\lg\frac{1+r}{1-r}$$

Розподіл величини z є майже незмінним за формою, так як мало залежить від чисельності вибірки і від значення коефіцієнта кореляції в генеральній сукупності. Величина змінює своє значення від $-\infty$ до $+\infty$, а її розподіл швидко наближається до нормального розподілу із середнім значенням

$$\overline{z} = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$$

і дисперсією

$$\sigma_z^2 = \frac{1}{n-3}$$

Перетворення коефіцієнта кореляції у величину z проводиться за запропонованою встановленої таблиці. У таблиці містяться величини z, що відповідають значенням емпіричного коефіцієнта кореляції r. Критерієм достовірності показника z служить таке ставлення:

$$t_z = \frac{z}{S_z} = z\sqrt{n-3}\ . \qquad (7.9)$$

Цей критерій придатний для вибірки будь-якого обсягу: вона використовується в усіх випадках, коли замість коефіцієнта кореляції r береться відповідає йому значення z. Нульова гіпотеза перевіряється за допомогою $t-$критерію Стьюдента для прийнятого рівня значущості та кількості степенів вільності $k = n - 2$.

Застосування z-перетворення дозволяє з більшою впевненістю оцінювати значимість емпіричного коефіцієнта кореляції, а, також і різниця між двома вибірковими коефіцієнтами $r_1 - r_2$, коли виникає такого роду необхідність.

Статистична недостовірність обчисленого на нечисленною вибірці коефіцієнта кореляції нічого, власне, не доводить. Адже при повторній вибірці нульова гіпотеза може виявитися неспроможною. Можна розрахувати необхідний обсяг вибірки для заданої точності коефіцієнта кореляції. Для цього служить формула

$$n = \frac{t^2}{z^2} + 3 \ , \tag{7.10}$$

де $n-$шуканий обсяг вибірки, $t-$ величина, задана за прийнятим рівнем значущості, z-перетворення (за Фішером) величина емпіричного коефіцієнта кореляції.

7.5 Оцінка різниці між коефіцієнтами кореляції

При порівнянні коефіцієнтів кореляції, обчислених на незалежних вибірках, нульова гіпотеза зводиться до припущення, що в генеральній сукупності різниця між цими показниками дорівнює нулю. Нульова гіпотеза перевіряється за допомогою t-критерію, який представляє відношення різниці між емпіричними коефіцієнтами кореляції r_1 і r_2 до її статистичної помилку, яка визначається за формулою

$$S_{dz} = \sqrt{S_{r_1}^2 + S_{r_2}^2} \ , \tag{7.11}$$

де $S_{r_1}^2$ і $S_{r_2}^2$ – помилки порівнюваних коефіцієнтів кореляції, які обчислюють за формулами (7.6) і (7.7), дивлячись за обсягами вибірок, для яких обчислені коефіцієнти кореляції.

Нульова гіпотеза відкидається при $t_{\hat{o}} \geq t_{st}$ для прийнятого рівня значущості (α) і числа степенів вільності $k = (n_1 - 2) + (n_2 - 2) = n_1 + n_2 - 4$.

t-критерій визначається за різницею $z_1 - z_2$, віднесеної до своєї помилку:

$$t_{dz} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}, \qquad (7.12)$$

тому що більш точна оцінка різниці між коефіцієнтами кореляції, обчисленими на нечисленних вибірках, виходить при використанні методу z, тобто на основі перетворених коефіцієнтів кореляції.

7.6 Кореляційне відношення

Вимірювання нелінійної залежності між варіюючими ознаками здійснюється за допомогою показника, запропонованого К. Пірсоном і названого **кореляційним відношенням**. На відміну від коефіцієнта кореляції, що характеризує залежність між випадковими змінними величинами Y і X з точки зору прямої пропорційності, кореляційне відношення, що позначається грецькою літерою η, описує її двосторонність.

Як і коефіцієнт кореляції, кореляційне відношення – величина відносна. Але на відміну від коефіцієнта кореляції кореляційне відношення завжди є величиною додатною, здатної приймати значення від 0 до 1. Коефіцієнт кореляції є рівнозначною заходом для обох кореляційно пов'язаних ознак Y і X, а коефіцієнти кореляційного відносини зазвичай не дорівнюють один одному, тобто $\eta_{xy} \neq \eta_{yx}$. Рівність між цими коефіцієнтами можливо тільки при строго лінійної залежності між ознаками. Кореляційне відношення є універсальним показником, тому що дозволяє характеризувати будь-яку форму кореляційного зв'язку.

Основний спосіб обчислення кореляційного відношення.

Коефіцієнти кореляційного відношення виражаються наступними формулами:

$$\eta_{yx} = \frac{S_{yx}}{S_y} \quad \text{і} \quad \eta_{xy} = \frac{S_{xy}}{S_x}, \qquad (7.13)$$

де S_y і S_x – загальні, а S_{yx} і S_{xy} – групові середні квадратичні відхилення.

S_{yx} і S_{xy} визначаються за формулами:

$$S_{yx} = \sqrt{\frac{1}{n}\sum p_x(\overline{y}_x - \overline{y})^2} \quad \text{і} \quad S_{xy} = \sqrt{\frac{1}{n}\sum p_y(\overline{x}_y - \overline{x})^2}, \qquad (7.14)$$

де $\overline{y}$ і $\overline{x}$ – загальні, а $\overline{y}_x$ і $\overline{x}_y$ – групові середні арифметичні димерної сукупності; p_y – частоти ряду Y; p_x – частоти ряду X; n – загальна кількість спостережень, або обсяг вибірки.

Питання для самоперевірки:

1. Для чого необхідний кореляційний аналіз?
2. Що таке аргумент?
3. Що таке функція?
4. Що називається функціональною залежністю?
5. Що називається кореляцією?
6. Назвіть завдання кореляційного аналізу.
7. Що таке коефіцієнт кореляції?
8. Що таке показник коваріації випадкової величини?
9. Охарактеризуйте емпіричний коефіцієнт кореляції.
10. Охарактеризуйте оцінку достовірності коефіцієнта кореляції.
11. Що таке число степенів вільності?
12. Охарактеризуйте z-перетворення Фішера
13. Що таке оцінка різниці між коефіцієнтом кореляції випадкової величини?
14. Що являють собою помилки коефіцієнтів кореляції, що порівнюються?
15. Охарактеризуйте кореляційне відношення.
16. Опишіть основний спосіб обчислення кореляційного відношення.

8 РЕГРЕСІЙНИЙ АНАЛІЗ

Термін «регресія» (від лат. *Regressio* – рух назад) ввів Гальтон. Вивчаючи статистичним методом успадкування кількісних ознак, він виявив, що потомство високорослих і низькорослих батьків відхиляється (регресує) від них на 1/3 у бік середнього рівня цієї ознаки в даній популяції.

Кореляційну залежність між ознаками можна описувати різними способами. Зокрема, будь-яка форма зв'язку може бути виражена рівнянням загального вигляду $y = f(x)$, де ознака y – **залежна змінна**, або **функція від незалежної змінної** X, яка називається **аргументом.** Відповідність між аргументом і функцією може бути задано таблицею, формулою, графіком і т.д. Весь цей арсенал засобів, що використовуються для опису кореляційних зв'язків, становить зміст регресійного аналізу. Зміна функції залежно від змін одного чи кількох аргументів, як уже повідомлялося, називається **регресією**.

Для вираження регресії служать **емпіричні** і **теоретичні ряди**, їхні графіки – **лінії регресії**, а також **кореляційні рівняння (рівняння регресії)** і **коефіцієнт лінійної регресії**.

Показники регресії виражають кореляційний зв'язок двосторонньо, враховуючи зміну середньої величини $\overline{y}_x$, ознаки Y при зміні значень x_i, ознаки X, і, навпаки, показують зміну середньої величини $\overline{x}_y$ ознаки X за зміненими значеннями y_i, ознаки Y. Виняток становлять тимчасові ряди, або ряди динаміки, що показують зміна ознак у часі. Регресія таких рядів є односторонньою.

Ряди регресії, особливо їх графіки, дають наочне уявлення про форму і тісноті кореляційного зв'язку між ознаками, в чому і полягає їх цінність. Форма зв'язку між біологічними ознаками може бути різноманітною. Завдання полягає у тому, щоб будь-яку форму кореляційного зв'язку виразити рівнянням певної функції (лінійна, параболічна), що дозволяє отримувати потрібну інформацію про кореляцію між змінними величинами Y і X, передбачати можливі зміни ознаки Y на основі відомих змін X, пов'язаного з Y кореляційно.

8.1 Рівняння лінійної регресії

Зазвичай ознака Y розглядається як функція багатьох аргументів – $x_1, x_2, x_3, \ldots, x_m$ – і може бути записана у вигляді

$$y = a + bx_1 + cx_2 + dx_3 + \ldots , \qquad (8.1)$$

де a, b, c і d – параметри рівняння, що визначають співвідношення між аргументами і функцією. У практиці враховуються не всі, а лише деякі аргументи, в простому випадку, як при описі лінійної регресії, – всього один:

$$y = a + bx \ . \qquad (8.2)$$

У цьому рівнянні параметр a – вільний член; графічно він представляє відрізок ординати (y) у системі прямокутних координат. Параметр b називається **коефіцієнтом регресії**. З точки зору аналітичної геометрії b – кутовий коефіцієнт, що визначає нахил лінії регресії по відношенню до осей координат. В області регресійного аналізу це параметр показує, наскільки в середньому величина однієї ознаки (Y) змінюється при зміні на одиницю заходи іншого кореляційно пов'язаного з Y ознаки X.

Наочне уявлення про те параметрі і про стан ліній регресії Y по X і X по Y в системі прямокутних координат дає рис.8.1.

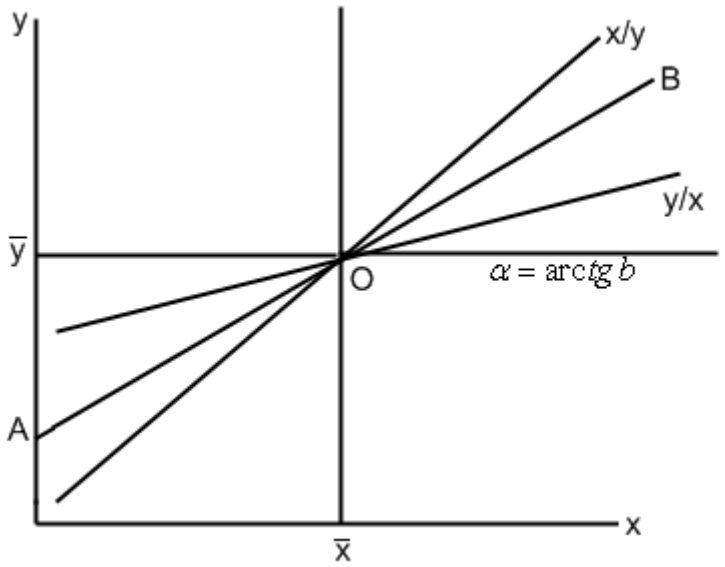


Рисунок – 8.1 Схема ліній регресії Y по X і X по Y в системі прямокутних координат.

Лінії регресії, як показано, перетинаються в точці $O(\bar{x},\bar{y})$, що відповідає середнім арифметичним значенням кореляційно пов'язаних один з одним ознак Y і X. Лінія AB, що проходить через цю точку, зображують повну (функціональну) залежність між, змінними величинами Y і X, коли коефіцієнт кореляції r=1. Чим сильніше зв'язок між Y і X,

тим ближче лінії регресії до AB, і, навпаки, чим слабкіший зв'язок між варіюючими ознаками, тим більш віддаленими виявляються лінії регресії від AB. За відсутності зв'язку між ознаками, коли $r = 0$, лінії регресії виявляються по прямим кутом (90°) по відношенню один до одного.

Коефіцієнт регресії показує, наскільки в середньому величина однієї ознаки (Y) змінюється при зміні на одиницю заходи іншого ознаки (X), пов'язаного з Y кореляційно. Оскільки дослідника може цікавити не тільки регресія Y по X, але і X по Y, то і коефіцієнт регресії відповідно позначається символами b_{yx} і b_{xy}. Ці показники визначаються за формулами

$$b_{yx} = r\frac{s_{y^i y}}{s_{x^i x}} \qquad \text{і} \qquad b_{xy} = r\frac{s_{x^i x}}{s_{y^i y}}. \tag{8.3}$$

За першою формулою визначається при зміні $\bar{y}_x$ на одиницю міри ознаки X, а по другій – значення $\bar{x}_y$ при зміні на одиницю міри Y.

В області кореляційно-регресійного аналізу велике значення має **коефіцієнт лінійної регресії**. Коефіцієнт регресії можна обчислити, минаючи розрахунок середніх квадратичних відхилень S_y і S_x:

$$b_{yx} = r\sqrt{\frac{\sum(y_i - \bar{y})^2}{\sum(x_i - \bar{x})^2}} \tag{8.4}$$

та

$$b_{xy} = r\sqrt{\frac{\sum(y_i - \bar{y})^2}{\sum(x_i - \bar{x})^2}}\ . \tag{8.5}$$

Якщо ж коефіцієнт кореляції невідомий, коефіцієнт регресії визначається за формулою

$$b_{yx} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \tag{8.6}$$

та

$$b_{xy} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(y_i - \bar{y})^2}\ . \tag{8.7}$$

для великих вибірок, згрупованих в варіаційні ряди,

$$b_{yx} = \frac{\sum p_{xy} xy - \frac{(\sum p_x x)(\sum p_y y)}{n}}{\sum p_x x^2 - \frac{(\sum p_x x)^2}{n}} \qquad (8.8)$$

та

$$b_{xy} = \frac{\sum p_{xy} xy - \frac{(\sum p_x x)(\sum p_y y)}{n}}{\sum p_y y^2 - \frac{(\sum p_y y)^2}{n}} \; . \qquad (8.9)$$

Зв'язок між коефіцієнтами регресії і кореляції. Порівнюючи формулу коефіцієнта регресії з основною формулою коефіцієнта кореляції, бачимо, що їх чисельники дорівнюють $\sum(y_i - \overline{y})(x_i - \overline{x})$. Це свідчить про певний зв'язок між цими характеристиками. Вона виражається рівністю $r^2 = b_{yx} b_{xy}$, звідки

$$r = \sqrt{b_{yx} b_{xy}} \; . \qquad (8.10)$$

Коефіцієнт кореляції дорівнює середньої геометричної з коефіцієнтом регресії. Ця формула цінна тим, що по-перше, може бути використана для знаходження невідомої величини коефіцієнта кореляції за відомими значеннями коефіцієнта регресії b_{yx} і b_{xy}, а по-друге, дозволяє контролювати правильність розрахунку коефіцієнта кореляції, якщо відомі величини b_{yx} і b_{xy}.

Як і коефіцієнт кореляції, коефіцієнт регресії характеризує тільки лінійну зв'язок і супроводжується знаком плюс (+) при позитивній або знаком мінус (-) при негативній зв'язках.

8.2 Визначення параметрів лінійної регресії

Визначення параметрів лінійної регресії – одне із завдань регресійного аналізу. Вона вирішується способом найменших квадратів, заснованим на вимозі, щоб сума квадратів відхилень варіант від лінії регресії була найменшою. Цій вимозі задовольняє наступна система нормальних рівнянь:

$$an + b\sum x = \sum y,$$

$$a\sum x + b\sum x^2 = \sum xy.$$

З спільного вирішення цих рівнянь отримуємо формули, зручні для визначення параметрів a і b безпосередньо за значеннями ознак Y і X, не вдаючись до складання системи рівнянь.

Ряди регресії – це ряди усереднених значень ($\overline{y}_x$ і $\overline{x}_y$) варіюючих ознак Y і X, що відповідають значенням аргументів y_i і x_i. Тому емпіричні рівняння регресії слід записувати так:

$$\overline{y}_x = a_{yx} + b_{yx}x \qquad \text{і} \qquad \overline{x}_y = a_{xy} + b_{xy}y \tag{8.11}$$

Формули для визначення параметрів a і b приймають такі вирази:

$$a_{yx} = \frac{\sum y\sum x^2 - \sum x\sum xy}{n\sum x^2 - (\sum x)^2} \quad \text{і} \quad a_{xy} = \frac{\sum x\sum y^2 - \sum y\sum xy}{n\sum y^2 - (\sum y)^2}, \tag{8.12}$$

або

$$a_{yx} = \overline{y} - b_{yx}\overline{x} \quad \text{і} \quad a_{xy} = \overline{x} - b_{xy}\overline{y}, \tag{8.13}$$

$$b_{yx} = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2} \quad \text{і} \quad b_{xy} = \frac{n\sum xy - \sum x\sum y}{n\sum y^2 - (\sum y)^2}, \tag{8.14}$$

або

$$b_{yx} = \frac{\sum xy - n\overline{x}\,\overline{y}}{\sum x^2 - n\overline{x}^2} \quad \text{і} \quad b_{xy} = \frac{\sum xy - n\overline{x}\,\overline{y}}{\sum y^2 - n\overline{y}^2}, \tag{8.15}$$

або

$$b_{yx} = \frac{\sum xy - \dfrac{\sum x\sum y}{n}}{\sum x^2 - \dfrac{(\sum x)^2}{n}} \quad \text{і} \quad b_{xy} = \frac{\sum xy - \dfrac{\sum x\sum y}{n}}{\sum y^2 - \dfrac{(\sum y)^2}{n}}. \tag{8.16}$$

Рівняння лінійної регресії можна виразити у вигляді відхилень варіант від їх середніх арифметичних:

$$\overline{y}_x - \overline{y} = b_{yx}(x_i - \overline{x}) \quad \text{і} \quad \overline{x}_y - \overline{x} = b_{xy}(\overline{y}_i - \overline{y}) \tag{8.17}$$

У такому випадку система нормальних рівнянь для визначення параметрів a і b буде наступна:

$$an + b\sum(x_i - \overline{x}) = \sum(y_i - \overline{y});$$

$$a\sum(x_i - \overline{x}) + b\sum(x_i - \overline{x})^2 = \sum(y_i - \overline{y})(x_i - \overline{x}).$$

Оскільки $\sum(x_i - \overline{x}) = 0$ і $\sum(y_i - \overline{y}) = 0$, то параметр b виразиться у вигляді наведених формул (8.6-8.7); параметр a легко знайти за формулою (8.13). Якщо середні $\overline{y}$ і $\overline{x}$ перенести в праву частину рівняння (8.17), отримаємо

$$\overline{y}_x = \overline{y} + b_{yx}(x_i - \overline{x}) \quad \text{і} \quad \overline{x}_y = \overline{x} + b_{xy}(y_i - \overline{y}) \tag{8.18}$$

де $a = \overline{y}$. Система нормальних рівнянь для визначення параметрів виявляється такою:

$$an + b\sum(x_i - \overline{x}) = \sum(y_i - \overline{y});$$

$$a\sum(x_i - \overline{x}) + b\sum(x_i - \overline{x})^2 = \sum(y_i - \overline{y})(x_i - \overline{x}).$$

Так як $\sum(x_i - \overline{x}) = 0$, система приймає такий вираз:

$$an = \sum y;$$

$$b\sum(x_i - \overline{x})^2 = \sum y(x_i - \overline{x}).$$

Звідси параметри рівняння лінійної регресії, вираженого у вигляді відхилень членів ряду від їх середньої величини, виразяться формулами

$$a = \frac{\sum y}{n}, \tag{8.19}$$

$$b = \frac{\sum y(x_i - \overline{x})}{\sum(x_i - \overline{x})^2}. \tag{8.20}$$

Ці формули зручні для визначення параметрів при знаходженні емпіричних рівнянь рядів динаміки.

8.3 Побудова емпіричних рядів регресії

Обчислені за значеннями однієї варіюючої ознаки X середні значення ($\overline{y}_x$) іншої ознаки Y, пов'язаного з кореляційно X, утворюють емпіричний ряд регресії. Це подвійний ряд чисел, який можна зобразити у системі прямокутних координат у вигляді емпіричної лінії регресії. При цьому по осі абсцис відкладаються значення незалежної змінної, або аргументу (X), а по осі ординат – середні значення залежної змінної ($\overline{y}_x$). Лінії регресії Y по X або X по Y дають наочне уявлення про форму і тісноті кореляційного зв'язку між варіюючими ознаками Y і X.

Методика побудови емпіричних рядів регресії досить проста і зводиться до знаходження групових середніх $\overline{y}_x$ і $\overline{x}_y$, що відповідають значенням ознак Y і X.

Емпіричні лінії регресії – ламані лінії в системі прямокутних координат. Пояснюється це тим, що поряд з основною тенденцією, що визначає головні напрямки ліній регресії, на них позначаються впливу численних другорядних (випадкових) причин, що порушують плавний хід регресії. У силу цього емпіричні лінії регресії зазвичай не бувають добре доступним для огляду і потребують вирівнюванні.

Способи вирівнювання емпіричних рядів регресії.

Заміна ламаних ліній регресії на плавно йдуть в системі прямокутних координат називається **вирівнюванням емпіричних ліній регресії**. Найбільш простий спосіб, що не вимагає обчислювальної роботи, – графічний. Сутність його полягає в наступному. Емпіричний ряд регресії зображується у вигляді лінійного графіка в системі прямокутних координат. Потім навіч визначаються серединні точки регресії, по яких за допомогою лінійки або лекала проводиться суцільна лінія. Недолік цього способу очевидна: він не виключає вплив на результати вирівнювання регресії індивідуальних властивостей дослідника. Тому в тих випадках, коли необхідна більш висока точність при заміні ламаних ліній регресії на плавно йдуть, користуються іншими способами вирівнювання емпіричних рядів.

Інший спосіб – спосіб ковзної середньої – зводиться до послідовного обчислення ряду середніх арифметичних з двох або трьох сусідніх членів емпіричного ряду регресії.

Він зручний особливо в тих випадках, коли емпіричний ряд представлений досить великою кількістю членів, так що втрата двох з них (крайніх), що неминуче при цьому способі вирівнювання, помітно не відіб'ється на його структурі.

Найбільш точний з усіх способів вирівнювання емпіричних рядів – **спосіб найменших квадратів**, запропонований в 1806 р. К. Гауссом і

незалежно від нього А. Лежандром. В основу його покладено теорема, згідно з якою сума квадратів відхилень варіант (x_i) від середньої арифметичної ($\overline{x}$) є величина найменша, тобто $\sum(x_i - \overline{x})^2 = \min$. Звідси і назва методу, який знайшов широке застосування не тільки в біології, але і в техніці. Метод найменших квадратів універсальний і застосовується в різних випадках при знаходженні кореляційних рівнянь та визначенні їх параметрів.

У загальному вигляді застосування його до вирішення конкретних завдань зводиться до наступного:

1. За геометричному місцем точок двох змінних Y і X в системі прямокутних координат підбирають математичне рівняння, найкращим чином відображає форму зв'язку між цими величинами.
2. Підставляючи в шукане рівняння відповідні емпіричні дані, утворюють систему нормальних рівнянь, розв'язуючи яку знаходять параметри цього рівняння.
3. Підставляючи параметри в загальне рівняння, отримують емпіричне рівняння регресії, що виражає залежність між змінними величинами Y і X.
4. Підставляючи в емпіричне рівняння значення однієї змінної (X), знаходять відповідні середні значення іншої змінної ($\overline{y}_x$). Так виходить ряд теоретично очікуваних (вирівняних) значень функції з відомим значенням аргументу.

Ряди динаміки та їх вирівнювання. Коли зміни ознаки розглядаються в залежності від фактора часу, регресія набуває односторонній характер, так як фактор часу не залежить від змін ознаки. Такі ряди регресії називаються **тимчасовими рядами** або **рядами динаміки**. Характерна особливість рядів динаміки – та, що в якості незалежної змінної тут завжди виступає **фактор часу**.

Емпіричні ряди динаміки вирівнюються будь-яким описаним вище способом, з яких кращий спосіб найменших квадратів. Якщо між змінними Y і X ряду динаміки існує лінійний зв'язок, зручною формою її вираження може служити рівняння (8.18), з якого випливає, що для визначення параметрів a і b при знаходженні емпіричних рівнянь рядів динаміки слугують формули (8.19) і (8.20).

Вирівнювати ряди динаміки по емпіричному рівнянню лінійної регресії можна спрощеним способом, беручи відхилення членів ряду незалежної змінної (X) не від середньої арифметичної, як у попередньому випадку, а від центральної (нульовий) точки ряду у вигляді натурального порядку чисел (1, 2, 3, 4, ...), коли ряд складається з непарної кількості

рівно віддалених один від одного членів (обов'язково з урахуванням знаків), а при наявності парного числа членів – через два інтервали, тобто 1, 3, 5, 7, ... (теж з урахуванням знаків). В обох випадках відхилення в бік більших значень членів ряду незалежної змінної беруться зі знаком плюс, а в бік менших значень – зі знаком мінус, так що в сумі всі відхилення від нульової точки ряду будуть дорівнювати нулю. Параметр a лінійної регресії визначається за формулою (8.19), тобто він дорівнює середньої арифметичної ($\bar{y} = \frac{\sum y}{n}$) з числових значень залежно змінної (Y), а параметр b знаходять за формулою

$$b = \frac{\sum xy}{\sum x^2}\ , \tag{8.21}$$

де b – відхилення членів ряду X від його серединної точки, де $x = 0$.

Параметр a і b рівняння лінійної регресії визначаються шляхом вирішення спрощеної системи нормальних рівнянь, коли $\sum x = 0$ і система набуває вигляду:

$$an = \sum y, \quad b\sum x^2 = \sum xy\ ,$$ звідки $a = \sum y / n$ і $b = \sum xy / \sum x^2$.

Величина $\sum x^2 = \frac{(n-1)n(n+1)}{3}$ при парному числі рівно віддалених один від одного членів ряду і $\sum x^2 = \frac{(n-1)n(n+1)}{12}$ при непарному числі членів ряду.

Звідси параметр b рівняння регресії визначається за допомогою таких формул:

$$b = \frac{3\sum xy}{(n-1)n(n+1)} \tag{8.22}$$

за наявності парного числа членів ряду,

$$b = \frac{12\sum xy}{(n-1)n(n+1)} \tag{8.23}$$

за наявності непарного числа членів ряду.

У цих формулах $n-$ число членів ряду динаміки.

Оцінка достовірності вибіркових показників регресії. Емпіричні показники регресії є оцінками відповідних генеральних параметрів. І так, як випадкові величини, вони супроводжуються статистичними помилками. Так, помилка вибіркового коефіцієнта регресії виражається в наступному вигляді:

$$S_{byx} = \sqrt{\frac{\sum(y_i - \bar{y})^2 - \frac{\left[\sum(y_i - \bar{y})(x_i - \bar{x})\right]^2}{\sum(x_i - \bar{x})^2}}{(n-2)\sum(x_i - \bar{x})^2}}$$

та

$$S_{bxy} = \sqrt{\frac{\sum(x_i - \bar{x})^2 - \frac{\left[\sum(y_i - \bar{y})(x_i - \bar{x})\right]^2}{\sum(y_i - \bar{y})^2}}{(n-2)\sum(y_i - \bar{y})^2}}. \qquad (8.24)$$

Якщо відомі r_{xy}, S_y і S_x помилку коефіцієнта регресії можна визначити за формулою

$$S_{byx} = \frac{s_y i_y}{s_x i_x}\sqrt{\frac{1-r^2}{n-2}} \quad \text{і} \quad S_{bxy} = \frac{s_x i_x}{s_y i_y}\sqrt{\frac{1-r^2}{n-2}}. \qquad (8.25)$$

Достовірність вибіркового коефіцієнта регресії перевіряється за допомогою t-критерію з $k = n-2$ числами степенів вільності і прийнятим рівнем значущості (α). Нульова гіпотеза зводиться до припущення, що в генеральній сукупності коефіцієнт регресії дорівнює нулю.

Статистичні помилки, якими супроводжуються емпіричні рівняння лінійної регресії, що позначаються символами S_{yx} и S_{xy} визначаються за формулами

$$S_{yx} = S_y\sqrt{1-r^2}\ ; \quad S_{xy} = S_x\sqrt{1-r^2}\ ; \qquad (8.26)$$

а також

$$S_{yx} = \sqrt{\frac{\sum(y_i - \bar{y}_x)^2}{n-2}} \quad \text{і} \quad S_{xy} = \sqrt{\frac{\sum(x_i - \bar{x}_y)^2}{n-2}} \qquad (8.27)$$

де S_y і Sx – середні квадратичні (відхилення рядів Y і X, $\bar{y}_x$ і $\bar{x}_y$ – середні значення залежних змінних Y і X, розраховані за рівнянням регресії.

Значення статистичної помилки лінії регресії полягає в тому, що вона вказує на величину можливих відхилень емпірично знайдених значень змінних Y і X від лінії регресії, побудованої за значеннями $\overline{y}_x$ або $\overline{x}_y$. Помилка лінійної регресії дозволяє з тією чи іншою ймовірністю встановлювати границі, що включають певну частку всіх емпірично знайдених значень змінних величин. Для цього служать рівняння

$$\overline{y}_x = (a + b_{yx}x) \pm ts_{yx},$$

$$\overline{x}_y = (a + b_{xy}y) \pm ts_{xy}.$$

Аналіз емпіричних регресій має велике прикладне значення. По рівнянню регресії можна оцінити, наприклад, фізичний розвиток окремо взятого виду риби по відношенню до прийнятої нормі, дати групову оцінку тій чи іншій популяції при відомих, заздалегідь встановлених, нормативах для генеральної сукупності і т. д. Так, за наявності кореляції між ознаками Y і X в інтервалі $\overline{y}_x \pm S_{yx}$ або $\overline{x}_y \pm S_{xy}$ перебуває близько 68% всіх варіант сукупності, розподіляється по нормальному закону. Якщо варіанти, що знаходяться в цьому інтервалі, вважати «нормальними», то віддалені нижче або вище меж цього інтервалу можна розглядати як виходять за границі прийнятої «норми».

У тих випадках, коли оцінці піддаються не окремі види, а цілі або вибіркові групи з їх середніми характеристиками, границі довірчого інтервалу лінійної регресії встановлюються за рівнянням

$$\overline{y}_x = (a + b_{yx}x) \pm \frac{ts_{yx}}{\sqrt{n}},$$

де n – чисельності, або обсяг, вибіркової групи.

8.4 Вирази регресії іншими рівняннями

Регресія, що виражається рівнянням показового типу. У тих випадках, коли основна тенденція емпіричного ряду регресії випливає або виявляється близькою до закону геометричної прогресії, його вдається описати рівнянням показникової або експоненціальної функції:

$$y = ab^x \quad \text{і} \quad y = ae^{xb}. \tag{8.28}$$

Використання рівнянь такого виду пов'язана з їх логарифмування, що дозволяє трансформувати ці рівняння в рівняння прямої лінії. У даному випадку маємо

$$\lg y = \lg a + x \lg b \quad . \tag{8.29}$$

Логарифмічний перетворення вихідного рівняння регресії не тільки полегшує обчислення параметрів a і b, але і служить свого роду контролем того, наскільки правильно обрано застосовується рівняння. Зокрема, про правильність вибору рівняння показникової функції можна судити у випадку, якщо точки x і $\lg y$ розташовуються в системі координат на одній прямій.

Визначення параметрів a і b рівняння (8.29) задовольняє система нормальних рівнянь

$$n \lg a + \lg b \sum x = \sum \lg y ;$$

$$\lg a \sum x + \lg b \sum x^2 = \sum (x \lg y) .$$

Спільне рішення цієї системи призводить до наступних формулах:

$$\lg a = \frac{1}{D}\left[\lg y \sum x^2 - \sum \left(x \lg y \right) - \sum x \right]$$

та

$$\lg a = \frac{1}{D}\left[\lg y \sum x^2 - \sum \left(x \lg y \right) - \sum x \right]$$

де $D = n \sum x^2 - (\sum x)^2$.

Із системи рівнянь і наведених формул випливає, що для відшукання параметрів a і b потрібно попередньо знайти величини $\sum x, \sum x^2, \sum \lg y, \sum (x \lg y).$

Розрахунки цих велич спрощуються, якщо члени ряду незалежної змінної X висловити числами натурального ряду 1, 2, 3, 4, ...

Регресія, що виражається рівнянням степеневого типу. Залежність між змінними величинами Y і X може виражатися рівнянням степеневої функції

$$y = ax^b \quad , \tag{8.30}$$

яке логарифмування перетворюється на рівняння прямої лінії

$$\lg y = \lg a + b \lg x \ . \tag{8.31}$$

Умовою правильності застосування цього рівняння служить вимога, щоб точки $\lg y$ і $\lg x$ в системі прямокутних координат знаходилися на одній прямій. Ця особливість відрізняє рівняння степеневої функції від рівняння функції показникової, коли в системі координат на одній прямій виявляються точки $\lg y$ і x.

Для визначення параметрів a і b рівняння степеневої функції служить наступна система нормальних рівнянь:

$$n \lg a + b \sum \lg x = \sum \lg y ;$$
$$\lg a \sum \lg x + b \sum (\lg x)^2 = \sum (\lg x \lg y) .$$

З спільного вирішення цієї системи виходять формули

$$\lg a = \frac{1}{D}\left[\sum \lg y \sum (\lg x)^2 - \sum (\lg x \lg y) \sum \lg x \right]$$

та

$$b = \frac{1}{D}\left[n \sum (\lg x \lg y) - \sum \lg x \sum \lg y \right],$$

де $D = n \sum (\lg x)^2 - (\sum \lg x)^2$.

З цих формул випливає, що для знаходження параметрів a і b потрібно попередньо розрахувати $\sum \lg y, \sum (\lg y \lg x), \sum \lg x, \sum (\lg x)^2$.

Регресія, що виражається рівнянням логістичної кривої. Серед численних форм кореляційної залежності між змінними величинами Y і X для біолога певний інтерес становить так звана логістична залежність, зображувана графічно у вигляді *S*-подібної кривої і описувана аналітичним рівнянням Ферхюльста

$$y = \frac{N}{1+10^{a+bt}} + C, \qquad (8.32)$$

де y – ознака, що враховується; t – час, що минув від початкової, чи базисною, величини ознаки (C), з якої розпочато його вимір та до граничної в даних умовах величини N, яку він досяг за час t, a і b - параметри рівняння, які визначають логістичність кривої.

Логістична закономірність спостерігається, наприклад, у зміні обсягу біомаси, а також щодо зростання чисельності особин даної популяції, коли початкова величина (C) спочатку зростає дуже швидко, а потім при наявності обмежуючих умов темп зростання чисельності популяції швидко знижується і вона переходить в стан динамічної рівноваги. Шляхом логарифмічного перетворення рівняння (8.32) набуває такий вигляд:

$$\lg\left(\frac{N}{y-c}-1\right) = a + bt$$

Позначивши, через Z, отримуємо рівняння лінійної регресії:

$$\lg z = a + bt. \qquad (8.33)$$

Визначенню параметрів a і b це рівняння задовольняє наступна система нормальних рівнянь:

$$an - b\sum t = \sum \lg z;$$

$$a\sum t + b\sum t^2 = \sum (t \lg z)$$

Вирішуючи спільно цю систему відносно параметрів a і b отримуємо такі формули

$$a = \frac{1}{D}\left[\sum \lg z \sum t^2 - \sum t \sum (t \lg z)\right]$$

$$b = \frac{1}{D}\left[n\sum (t \lg z) - \sum t \sum \lg z\right]$$

де $D = n\sum t^2 - (\sum t)^2$

З цих формул випливає, що для отримання логістичної залежності між змінними і необхідно попередньо розрахувати $\sum t$, $\sum t^2$, $\sum \lg z$, $\sum (t \lg z)$;

потім, визначивши параметри a і b, знайти для кожного значення t (в межах обліковується відрізка часу) величини $\lg z$ і z, що і призведе в кінцевій інстанції до знаходження очікуваних значень $\overline{y}_t$.

Питання для самоперевірки:

1. Для чого необхідний регресійний аналіз?
2. Що таке регресія?
3. Що таке залежна змінна?
4. Що таке аргумент?
5. Які показники необхідні для вираження регресії?
6. Охарактеризуйте рівняння лінійної регресії.
7. Що таке коефіцієнт регресії і що він показує?
8. Охарактеризуйте коефіцієнт лінійної регресії.
9. Охарактеризуйте зв'язок між коефіцієнтом регресії і коефіцієнтом кореляції.
10. Як визначають параметри лінійної регресії.
11. Що таке ряди регресії?
12. Як відбувається побудова емпіричних рядів регресії.
13. Назвіть способи вирівнювання емпіричних рядів регресії.
14. Охарактеризуйте спосіб найменших квадратів.
15. Що таке ряди динаміки або тимчасові ряди?
16. Що таке фактор часу?
17. Охарактеризуйте спосіб вирівнювання рядів динаміки.
18. Охарактеризуйте рівняння регресії показникового типу.
19. Охарактеризуйте рівняння регресії степеневого типу.
20. Охарактеризуйте рівняння регресії логістичної кривої.

9 ДИСПЕРСІЙНИЙ АНАЛІЗ

Поряд з кореляційним і регресійним аналізом при вивченні причинно-наслідкових відносин між явищами особливо цінним виявився **метод дисперсійного аналізу** запропонований **Р.Е. Фішером** (1925) і вдосконалений його послідовниками. Цей метод заснований на розкладанні загальної дисперсії статистичного (дисперсійного) комплексу на складові компоненти, порівнюючи які один з одним за допомогою F – критерію, можна визначити частку загальної варіації досліджуваного (результативного) ознаки, обумовлену дією на нього як регульованих, так і не регульованих в досвіді факторів.

Дисперсійний аналіз як метод комплексної оцінки вибіркових показників пред'являє певні вимоги до групування вибіркових даних і планування спостережень. Результати спостережень підлягають дисперсійного аналізу, групуються з урахуванням підрозділів кожного регульованого фактора, що впливає на ознаку, наприклад, по дозам добрив термінами або способів внесення їх у грунт, приналежності тварин до тієї чи іншої породної чи племінний групі, їх віковим складом і т. д. Якщо регульований фактор впливає на ознаку, то воно неодмінно позначиться на величині групових середніх, які будуть істотно відрізнятися один від одного. Усередині кожної групи теж виявиться варіювання, викликане впливом на ознаку не регулюються в досвіді факторів. Залежність між цими джерелами варіювання виразиться рівністю

$$\sum_{i}^{N}(x_i-\bar{x})^2 = n\sum_{I}^{a}(\bar{x}_i-\bar{x})^2 + \sum_{I}^{a}\left[\sum_{I}^{n}(x_i-\bar{x}_i)^2\right], \qquad (9.1)$$

де перший член, що позначається надалі символом D_y – сума квадратів відхилень окремих варіант (x_i), або (за термінологією Фішера), всього комплексу спостережень від їх загальної середньої ($\bar{x}$). Другий член рівняння, що позначається нами символом, D_x – сума квадратів відхилень (a) групових чи приватних середніх ($\bar{x}_i$) від загальної середньої комплексу, помножена на число варіант у групах (n). Третій член рівняння, що позначається символом D_z – сума з сум квадратів відхилень окремих варіант від їх групових середніх.

Віднесенням сум квадратів відхилень до числах степенів вільності (k) виходять вибіркові дисперсії $S_y^2 = D_y / k_y$; $S_x^2 = D_x / k_x$ і $S_z^2 = D_z / k_z$ які

служать оцінками відповідних генеральних параметрів; S_y^2 – загальної дисперсії комплексу (σ_y^2), S_x^2 – міжгрупової дисперсії (σ_x^2) і S_z^2 – дисперсії внутрішньогрупової, чи залишкової (σ_z^2). Відношення дисперсії міжгрупової, або факторіальної, до дисперсії внутрішньогрупової, або залишкової $F = S_x^2 / S_z^2$, слугує критерієм оцінки впливу на ознаку регульованих у досліді факторів.

Нульова гіпотеза (H_0) зводиться до припущення, що генеральні міжгрупові середні рівні і дисперсії груп і міжгрупові генеральні дисперсії не розрізняються. Іншими словами, нульова гіпотеза виходить з того, що ніякого систематичного дії регульованих факторів на результативний ознака не існує і спостережувані між груповими середніми відмінності випадкові. Це припущення, або нульова гіпотеза, спростовується, якщо $F_{\hat{o}} \geq F_{st}$ для прийнятого рівня значущості (α) і чисел степенів вільності k_x і k_z. В іншому випадку, тобто коли $F_{\hat{o}} < F_{st}$, нульова гіпотеза зберігається і відмінності, що спостерігаються між груповими середніми комплексу, визнаються несуттєвими, випадковими. Після докази дії регульованого фактора або факторів, або їх спільної дії на ознаку статистичної достовірності, переходять, коли це необхідно, до порівняння групових середніх один з одним або з іншими показниками (загальною середньою комплексу, з прийнятою нормою, стандартом і т. д.) .

Заключний етап дисперсійного аналізу – оцінка сили впливу окремих факторів або їх спільної дії на результативну ознаку.

Дисперсійний аналіз характеризується *суворої логічністю і послідовністю обчислювальних операцій*. Цінність цього методу полягає в тому, що він дозволяє виявити сумарна дія факторів, дія кожного регульованого в досвіді чинника, а також дію різних поєднань факторів один з одним на результативний ознака.

Дисперсійний аналіз виник у процесі удосконалення методики сільськогосподарської дослідної справи, але незабаром знайшов широке застосування не тільки в біології і суміжних з нею галузях знання, але і в техніці, а також у педагогіці і психології при вирішенні багатьох комплексних завдань.

Правильне застосування дисперсійного аналізу передбачає нормальне або близьке до нормального розподіл сукупності, з якої взято вибірки, що об'єднуються в дисперсійний комплекс. При цьому важливо, щоб дисперсії вибіркових груп були однаковими або не дуже різнилися. При плануванні спостережень та обробці результатів слід також прагнути до того, щоб у

групах дисперсійного комплексу знаходилося однакове або пропорційну кількість варіант, що значно полегшує дисперсійний аналіз.

Ознаки, що змінюються під впливом тих чи інших причин, називаються **результативними**, а причини, що викликали зміну величини результативної ознаки або ознак, – **факторами**. Наприклад, маса або лінійні розміри, за якими судять про організм, його фізичний розвиток, і т. п. – все це ознаки, на які впливають різні чинники: елементи або режим харчування, дози лікарських або токсичних речовин і т. п. Фактори позначаються заголовними буквами латинського алфавіту $A, B, C, \ldots$, а обліковуються ознаки – через $X, Y, Z, \ldots$.

Факторів, що впливають на один і той самий ознака, багато. У досвіді ж регулюються лише деякі з них: вони називаються регульованими або організованими факторами на відміну від тих, які регулюванню не піддаються, хоча і впливають на величину результативної ознаки. Зазвичай кожен регульований фактор випробовується серійно, тобто у вигляді кількох відокремлених один від одного груп, званих градаціями. Їх прийнято позначати тими ж літерами, якими позначаються фактори. Наприклад, градації фактора A позначаються через A_1, A_2, A_3 і т. д., а фактора B – B_1, B_2, B_3 і т. д. Числа градацій того чи іншого чинника визначаються умовами досвіду, наприклад випробовуються дозами добрив, кількістю сортів при вивченні їх врожайності і т. д. Результативні ознаки теж можуть підрозділятися на окремі градації, на яких випробовується дію регульованих факторів.

Як вже згадувалося, дисперсійний аналіз дозволяє враховувати не тільки спільна дія регульованих факторів, але і дія кожного з них окремо, а також дію різних комбінацій цих факторів на результативний ознака. Цього, однак, не можна сказати про не регульованих в досвіді чинниках, дію яких на ознаку враховується не диференційовано, а сумарно. Нарешті, слід відзначити, що дисперсійний аналіз не виключає можливості висловлювати обліковуються ознаки не тільки в абсолютних одиницях виміру та рахунки, але і в балах, індекси та інших відносних і умовних одиницях.

Умови утворення і види статистичних комплексів. Статистичні, або дисперсійні, комплекси можуть формуватися як в плановому порядку, так і на основі вже зібраних вихідних даних, які піддаються дисперсійного аналізу. При утворенні будь-якого статистичного комплексу, а також і при плануванні експерименту необхідно дотримуватися принаймні дві основні умови: по-перше, враховуються фактори повинні бути незалежні один від одного, а по-друге, вибірка, згрупованих у статистичний комплекс, повинна формуватися за принципом рендомізації, тобто методом

випадкового відбору варіант з генеральної сукупності, розподіляється по нормальному закону. Структура комплексу визначається числом градацій регульованого фактора або факторів, а також числом підрозділів або груп, утворених за результативному ознакою. Форма дисперсійного комплексу задається таблицею, в якій число рядків дорівнює числу градацій результативної ознаки, а число стовпців відповідає числу градацій регульованого фактора або декільком факторів з їх градаціями.

Дисперсійний комплекс називається **однофакторним**, якщо випробовується дія на ознаку одного регульованого фактора, а якщо одночасно випробовується дія на ознаку двох, трьох або більшої кількості регульованих чинників, комплекс буде відповідно **дво-, три- і багатофакторним**. Результати масових випробувань, виражені у вигляді числових значень результативної ознаки, тобто маса варіантів, можуть розподілятися за градаціями комплексу *рівномірно і нерівномірно, пропорційно і непропорційно*. Звідси дисперсійні комплекси називаються **рівномірними, пропорційними і нерівномірними**. *Рівномірні та пропорційні комплекси* називаються **ортогональними**, а нерівномірні - **неортогональними**. У багатофакторних ортогональних комплексах виконується рівність між факторіальними сумами квадратів, тоді як у нерівномірних багатофакторних комплексах ця рівність порушується. Цю особливість слід враховувати під час планування дослідів, а під час проведення дисперсійного аналізу, прагнути до того, щоб у градаціях комплексу були однакові або пропорційні числа варіант, що значно спрощує обчислювальну роботу.

9.1 Аналіз однофакторних комплексів

9.1.1 Рівномірні комплекси

Однофакторні дисперсійні комплекси можуть бути рівномірними і нерівномірними. Незалежно від цього техніка дисперсійного аналізу однофакторних комплексів зводиться головним чином до розрахунку показників варіювання, якими в області дисперсійного аналізу служать середні квадрати відхилень, чи дисперсії, а також і до розрахунку групових чи приватних середніх ($\overline{x}_i$) і обший середньої арифметичної для всього комплексу в цілому ($\overline{x}$). Дисперсійний аналіз проводиться звичайно по тій чи іншій схемі. Для однофакторних рівномірних комплексів такою схемою може служив, наступний порядок операцій.

1. Вихідні дані групуються у вигляді комбінаційної таблиці таким чином, щоб градації регульованого фактора (A) розташовувалися по горизонталі у верхній частині таблиці, утворюючи її графи або стовпці, а значення результативної ознаки (X), тобто варіанти або дати (x_i), групувалися відповідно по градаціям фактора A.

2. Згрупувавши вибірку, як зазначено в п. 1, переходять до розрахунку допоміжних величин, потрібних для визначення сум квадратів відхилень $D_y = \sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum x_i^2 - \left[(\sum x_i)^2 \Big/ N \right]$ – загальна сума квадратів; міжгрупова сума квадратів розраховується наступним формулам:

$$D_A = \sum_{fi}^{a} \left[(\sum x_i)^2 \Big/ n - (\sum x_i)^2 \Big/ N \right]; \qquad (9.2)$$

$$D_A = n \sum_{fi}^{a} (\bar{x}_i - \bar{x})^2 \; ; \qquad (9.3)$$

або

$$D_A = n \sum_{i}^{a} \bar{x}_i^2 - \frac{(\sum x_i)^2}{N} \; . \qquad (9.4)$$

Внутрішньогрупову, або залишкову, суму квадратів можна визначити за різницею між загальною і міжгруповою сумами квадратів: $D_z = D_y - D_A$. Маючи на увазі $\bar{x} = \sum x_i / N$, звідки $\sum x_i = \bar{x}N$, величину $\left(\sum x_i\right)^2 / N$, яку будемо позначати надалі буквою H, можна виразити у вигляді $H = (\bar{x}N)^2 / N$.

Тут x_i – варіанти, що входять до складу всього комплексу і його окремих груп, або градацій; $\bar{x}$ загальна, а $\bar{x}_i$ – групові середні арифметичні; $n -$ число варіант в окремих групах (градаціях) комплексу; $N = \sum n$ – загальне число варіант, або обсяг дисперсійного комплексу.

3. Закінчивши розрахунок, визначають числа степенів вільності (k):

$k_y = N - 1$ – для загального варіювання;

$k_A = a - 1$ – для варіації міжгруповою (факторіальною);

$k_z = N - a$, або $k_z = (N-1) - (a-1)$, – для варіації внутрішньогрупової або залишкової. Тут a – число градацій фактора A.

Слід мати на увазі, що числа степенів вільності відповідно до рівністю $D_y = D_A + D_z$ знаходяться між собою в певних кількісних співвідношеннях: $k_y = k_A + k_z$. За цим рівності можна констатувати правильність розрахунку цих величин.

4. Визначення середніх квадратів відхилень або дисперсій з відносин сум квадратів відхилень до відповідних числах степенів вільності, тобто загальна дисперсія для всього комплексу дорівнює

$$S_A^2 = \frac{D_y}{N-1}$$

Міжгрупова дисперсія виражається формулою

$$S_A^2 = \frac{D_A}{a-1}$$

Внутрішньогрупова, або залишкова, дисперсія визначається за формулою $S_z^2 = \frac{D_z}{N-a}$

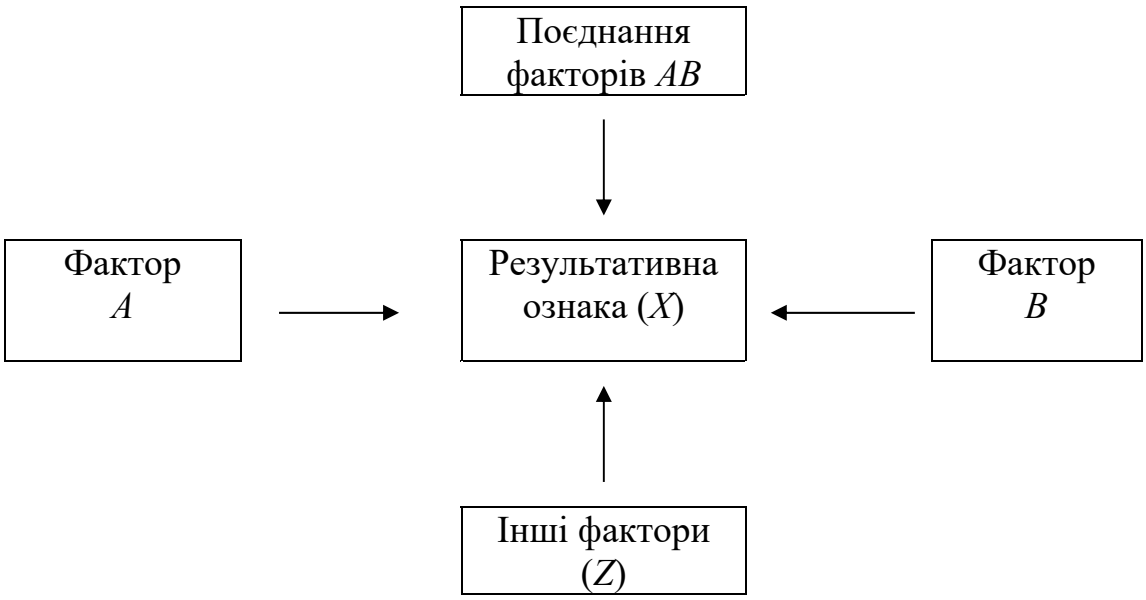
5. Нарешті, визначається ефективність дії фактора на результативний ознака. Для цього служить дисперсійне відношення, чи критерій Фішера $F = S_A^2 / S_z^2$ (при $S_A^2 \geq S_z^2$).

Так як дисперсійне відношення $F = S_A^2 / S_z^2$ – величина випадкова, його необхідно порівнювати з табличним (стандартним) значенням (F_{st}) критерію Фішера для прийнятого рівня значущості (α) і чисел степенів вільності k_A і k_z. При цьому, як зазначалося вище, кількість ступенів свободи для більшої дисперсії міститься по горизонталі, а для меншої дисперсії - у першому стовпчику таблиці Фішера. Нульова гіпотеза спростовується і ефективність дії фактора A на результативну ознаку визнається статистично достовірною, якщо $F\hat{o} \geq F_{st}$. В іншому разі нульова гіпотеза зберігається.

9.1.2 Аналіз двофакторних рівномірних комплексів

Переходячи до аналізу двофакторно рівномірних комплексів, що містять в градаціях факторів однакові числа варіант, слід зауважити, що

принципової різниці між аналізом багатофакторних комплексів і схемами, що застосовуються при аналізі однофакторних дисперсійних комплексів, немає. Багатофакторний аналіз не змінює, а лише трохи ускладнює загальну схему, оскільки поряд з дією кожного чинника в окремо доводиться враховувати і їхня спільна дія на результативний ознака. Так, якщо мається на увазі два регульованих фактора *A* і *B*, то їх вплив можна зобразити у вигляді такої схеми:



Тут загальна сума квадратів відхилень (D_y) містить не два, а чотири компоненти варіювання:

$$D_y = D_A + D_B + D_{AB} + D_Z \ ,$$

а факторіальна сума квадратів відхилень (Dx) складається з трьох компонентів:

$$Dx = D_A + D_B + D_{AB} \ .$$

Якщо ж враховуються не два чинники, а три регульовані фактора *A, B* і *C*, то поряд з їх індивідуальним дією можливо ще й дію на ознаку трьох попарних сполучень – *AB, AC* і *BC*, а також їх спільна дія *ABC* плюс вплив неорганізованих (випадкових) чинників. Таким чином, загальний компонент варіювання буде містити вісім елементів:

$$D_y = D_A + D_B + D_C + D_{AB} + D_{BC} + D_{ABC} + D_Z \ .$$

При більшій кількості чинників, що враховуються число їх можливих сполучень буде ще більше. У вивченні впливу на результативну ознаку

всіх чинників, що враховуються та їх можливих сполучень і полягає основна задача дисперсійного аналізу. При цьому не обов'язково у всіх випадках враховувати всі взаємодії регульованих факторів. Це питання вирішується дослідником в залежності від мети дослідження та повноти дисперсійного аналізу.

Дисперсійний аналіз двофакторну рівномірних (і пропорційних) комплексів проводиться за такою приблизною схемою.

1. Як і при обробці однофакторних комплексів, розраховується загальна сума квадратів відхилень:

$$D_y = \sum x_i^2 - H \ .$$

2. Потім визначаються загальна факторіальна (D_x) і залишкова (D_z) суми квадратів відхилень:

$$D_x = \sum \frac{(x_i)^2}{n} - H \quad \text{чи} \quad Dx = D_A + D_B + D_{AB}, \ D_z = D_y - D_x .$$

3. Далі визначаються сума квадратів відхилень для факторів A і B:

$$D_A = \sum \frac{(\sum x_A)^2}{n_A} - H \ , \text{або} \ D_A = \sum \frac{(\sum x_A)^2}{ab} - H ;$$

$$D_B = \sum \frac{(\sum x_B)^2}{n_B} - H \ , \text{або} \ D_B = \sum \frac{(\sum x_B)^2}{an} - H .$$

4. Нарешті, визначається сума квадратів відхилень для спільної дії факторів:

$$D_{AB} = Dx - (D_A + D_B) \ , \text{або} \ D_{AB} = Dx - D_A - D_B .$$

У цих формулах величина $H = \frac{(\sum x_i)^2}{N}$; x_i - варіанти, що входять до складу дисперсійного комплексу; $\sum x_A$ - сума варіант за фактором A; $\sum x_B$ - сума варіант за фактором B, $n_A = ab$ - кількість варіант в градаціях фактора A, $n_B = an$ - кількість варіант в градаціях фактора B; $n-$ чисельність варіант в окремих клітинах комбінаційної таблиці; $N = abn$ - загальна чисельність варіант, входять до складу дисперсійного комплексу; $a-$ число градацій фактора A; $b-$ число градацій фактора B.

5. Переходимо до встановлення чисел степенів вільності, які є однаковими для

- загальної дисперсії: $k_y = N - 1$,
- загальної факторіальною дисперсії: $k_x = ab - 1$,
- дисперсії за фактором: $k_A = a - 1$,
- дисперсії за фактором B: $k_B = b - 1$,
- дисперсії спільної дії: $k_{AB} = (a-1)(b-1)$,
- залишкової дисперсії: $k_Z = N - ab$.

При цьому, як уже зазначалося, числа степенів вільності повинні перебувати в таких же кількісних відносинах, як і відповідні їм суми квадратів відхилень:

$$D_y = D_A + D_B + D_{AB} + D_Z \quad \text{і} \quad k_y = k_A + k_B + k_{AB} + k_z;$$

$$Dx = D_A + D_B + D_{AB} \quad \text{і} \quad k_x = k_A + k_B + k_{AB};$$

$$D_y = D_x + D_z \quad \text{і} \quad k_y = k_x + k_z.$$

Наведені рівності можуть служити для перевірки правильності розрахунку сум квадратів відхилень і чисел степенів вільності.

6. Віднесенням сум квадратів відхилень до відповідних числах степенів вільності визначаються дисперсії, а за їх ставленням до величиною залишкової дисперсії визначається F-критерій, який порівнюється з критичним значенням (F_{st}) по таблиці Фішера для прийнятого рівня значущості та відповідних чисел степенів вільності (k). Нульова гіпотеза спростовується, якщо $F_{\hat{o}} \geq F_{st}$.

Заключним етапом дисперсійного аналізу є зведення результатів у таблицю.

У цій таблиці, як правило, наводяться і значення F_{st} для 5%-ного і 1%-ного рівнів значущості, що полегшує робити висновки щодо перевірки нульової гіпотези.

9.2 Аналіз двофакторних нерівномірних комплексів

Як вже повідомлялося, дисперсійні комплекси з неоднаковою, непропорційну чисельністю варіант в градаціях називаються **нерівномірними**, або **неортогональними комплексами**. Аналіз їх має свої особливості, пов'язані з диспропорційністю у розподілі варіант по

градаціях факторів A і B, що порушує рівності $Dx = D_A + D_B + D_{AB}$ і $D_{AB} = Dx - (D_A + D_B)$. Зберігається тільки рівність $D_y = D_x + D_z$. Тому при обробці таких комплексів обчислюються некореговані суми квадратів, що позначаються нами тими ж символами з додатком знака «штрих», тобто D'_A, D'_B, D'_{AB}, які потім піддаються виправленню, щоб замість $Dx \neq D_A + D_B + D_{AB}$ здійснювалося рівність $Dx = D_A + D_B + D_{AB}$, а отже, і рівність $D_{AB} = Dx - (D_A + D_B)$.

Виправлення некорегованих сум квадратів проводиться множенням їх на поправочний коефіцієнт $K = \frac{D_x}{D'_x}$, де

$$D'_x = N\left(\frac{\sum \bar{x}_i^2}{ab} - \bar{x}_0^2\right),$$

а невиправлені факторні суми квадратів (девіати) обчислюються за формулами:

$$D'_A = N\left(\frac{h_A^2}{a} - \bar{x}_0^2\right),$$

$$D'_B = N\left(\frac{h_B^2}{b} - \bar{x}_0^2\right)$$

$$D'_{AB} = D'x - (D'_A + D'_B)$$

У цих формулах $\bar{x}_0^2 = \left(\frac{\sum \bar{x}_i}{ab}\right)^2$;

$$h_A^2 = \sum\left(\frac{\sum \bar{x}_A}{b}\right)^2 \quad \text{та} \quad h_B^2 = \sum\left(\frac{\sum \bar{x}_B}{a}\right)^2$$

де $\bar{x}_i = \frac{x_i}{n}$ - групові середні; $\sum \bar{x}_A$ - сума групових для градацій A; $\sum \bar{x}_B$ - сума групових для градацій B; a - число градацій фактора A; b - число градацій фактора B (в групі A).

9.3 Аналіз ієрархічних комплексів

Поряд з розглянутими схемами, в яких можливі будь-які комбінації впливають на ознаку факторів, у практиці доводиться вдаватися до організації і таких дисперсійних комплексів, у яких вільне комбінування чинників один з одним виключено. Ці комплекси, звані ієрархічними, організовуються, наприклад, при вивченні впливу батьків на продуктивність або поведінку їх потомства, при з'ясуванні взаємовідносин між спорідненими в систематичному відношенні групами живих істот і в інших подібних випадках. Характерна особливість їх - певна ієрархічна підпорядкованість структурних компонент, коли групи відносно низького положення знаходяться в строгій залежності від пов'язаних з ними груп більш високого рангу.

Аналіз ієрархічних комплексів має свої особливості, обумовлені неможливістю вільного комбінування різних груп за фактором B з різних градацій фактора A, що займає більш високе становище в загальній схемі дисперсійного ієрархічного комплексу. При обробці таких комплексів не обчислюється дисперсія взаємодії (S_{AB}), трохи інакше виглядають дисперсійні відносини (F), по-іншому, ніж при обробці звичайних комплексів, визначаються факторіальних дисперсії.

Ієрархічні комплекси можуть бути рівномірними, пропорційними і нерівномірними. Структура ієрархічного комплексу залежить від кількості врахованих чинників і їх градацій. Найпростіша ієрархічна схема – двофакторна схема дисперсійного аналізу.

Питання для самоперевірки:

1. Для чого необхідний дисперсійний аналіз?
2. Як відбувається групування первинних даних і планування досліджень при дисперсійному аналізі.
3. Охарактеризуйте етапи дисперсійного аналізу.
4. Які ознаки називаються результативними?
5. Що називається чинниками.
6. Які умови необхідні для утворення дисперсійного аналізу?
7. Що таке дисперсійний комплекс?
8. Назвіть види статистичних комплексів.
9. Який дисперсійний комплекс називається однофакторним?

10. Який дисперсійний комплекс називається двофакторним?
11. Який дисперсійний комплекс називається трифакторним?
12. Який дисперсійний комплекс називається багатофакторним?
13. Який однофакторний дисперсійний комплекс називається рівномірним?
14. Який однофакторний дисперсійний комплекс називається нерівномірним?
15. Який однофакторний дисперсійний комплекс називається пропорційним?
16. Охарактеризуйте аналіз двофакторних рівномірних комплексів.
17. Що таке рівномірне комплекс?
18. Охарактеризуйте аналіз двофакторних нерівномірних комплексів.
19. Що таке нерівномірний комплекс?
20. Охарактеризуйте аналіз ієрархічних комплексів.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ ПОСИЛАНЬ

Основна

1. Крюкова М.І. Статистичні методи в біологічних дослідженнях: Конспект лекцій. Одеса, ОДЕКУ, 2012. 118 с.

2. Барановський Д.І. Біометрія в програмному середовищі MS Excel: навчальний посібник / Д. І. Барановський, О. М. Гетманець, А. М. Хохлов. – Х. : СПД Бровін О. В., 2017. 90 с.

3. Осадча Ю. В. Математичні методи в біології: навч. посіб. Київ: 2017, 601 с.

4. Близнюченко О.Г. Біометрія: монографія / О. Г. Близнюченко. Полтава.: Редакційно-видавничий відділ «Terra» Полтавської державної аграрної академії, 2003. 346 с.

5. Горошко М.П. Біометрія: навчальний посібник / Горошко М. П., Миклуш С.І., Хомюк П.Г. Львів: Камула, 2004. 236 с.

Додаткова

1. Біометрія навч. посіб. для студ. вищ. навч. закл.: у 2 ч. / Є. Я. Швець, М. Г. Сидоренко, І. В. Червоний ; Запорізька державна інженерна академія. Запоріжжя, 2004

2. Біометрія навч. посіб. для студ. вищих навч. закл. / М. П. Горошко [и др.] ; Український держ. лісотехнічний ун-т. Л.:Камула, 2004. 235 с.

3. www.library-odeku.16mb.com

4. eprints.library.odeku.edu.ua

ДОДАТКИ

Додаток А

Значення інтеграла ймовірності для різних t

t	Соті частки t									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,0	0000	0080	0160	0239	0319	0399	0478	0558	0638	0717
0,1	0797	0876	0955	1034	1114	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2288
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2961	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778
1,0	6827	6875	6923	6970	7017	7063	7109	7154	7199	7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	9620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7995	8030
1,3	8064	8098	8182	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8788	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9089
1,7	9108	9127	9146	9164	9182	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9425	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9585	9596	9608	9615	9625	9634
2,1	9643	9652	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9755	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9717	9822	9827	9832
2,4	9836	9840	9845	9849	9853	9857	9861	9866	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9909	9912	9915	9917	9920	9922	9924	9926	9929
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947
2,8	9949	9950	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9967	9969	9970	9971	9972
3,0	9973	9974	9975	9976	9976	9977	9978	9979	9979	9980
3,1	9981	9981	9982	9983	9983	9984	9984	9985	9985	9986
3,2	9986	9987	9987	9988	9988	9988	9982	9989	9990	9990
3,3	9990	-	9991	-	9992	-	9998	-	9993	-
3,4	9993	-	9994	-	9994	-	9995	-	9995	-
3,5	9995	-	9996	-	9996	-	9996	-	9997	-

Додаток Б

Значення функції $f(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$

t	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0.0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0.1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0.2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0.3	3814	3802	3700	3778	3765	3752	3739	3725	3712	3697
0.4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0.5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0.6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0.7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0.8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0.9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1.0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1.1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1.2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1.3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1.4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315

Продовження додатку Б

t	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1.5	0,1295	0,1276	0,1257	0,1238	0,1219	0,1200	0,1182	0,1163	0,1145	0,1127
1.6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1.7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1.8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1.9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2.0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2.1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2.2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2.3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2.4	0224	0219	0213	0203	0203	0198	0194	0189	0184	0180
2.5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2.6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2.7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2.8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2.9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3.0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034

Додаток В

Значення функції Пуассона $P_n(m) = \frac{a^m}{m\ !} e^{-a}$

m	a									
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0	0.9048	8187	7408	6703	6065	5488	4966	4493	4066	3679
1	0905	1637	2222	2681	3033	3293	3476	3595	3659	3679
2	0045	0164	0333	0536	0758	0988	1217	1438	1647	1839
3	0002	0011	0033	0072	0126	0198	0284	0383	0494	0613
4	0000	0001	0003	0007	0016	0030	0050	0077	0111	0253
5	-	-	-	0001	0002	0004	0007	0012	0020	0031
6	-	-	-	-	-	-	0001	0002	0003	0005
7	-	-	-	-	-	-	-	-	-	0001

Продовження додатку В

m	a									
	1.1	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**	**2.0**
0	0.3329	3012	2725	2466	2231	2019	1827	1653	1496	1353
1	3662	3614	3543	3452	3347	3230	3106	2975	2842	2707
2	2014	2169	2303	2417	2510	2584	2640	2678	2700	2707
3	0738	0867	0998	1128	1255	1378	1496	1607	1710	1805
4	0203	0260	0324	0395	0471	0551	0636	0723	0812	0902
5	0045	0063	0084	0111	0141	0176	0216	0260	0309	0361
6	0008	0013	0018	0026	0035	0047	0061	0078	0098	0120
7	0001	0002	0003	0005	0008	0011	0015	0020	0027	0034
8	-	-	0001	0001	0001	0002	0003	0005	0006	0009
9	-	-	-	-	-	-	0001	0001	0001	0002

Продовження додатку В

m	*a*									
	2.1	**2.2**	**2.3**	**2.4**	**2.5**	**2.6**	**2.7**	**2.8**	**2.9**	**3.0**
0	0.1225	1108	1003	0903	0821	0743	0672	0608	0550	0498
1	2572	2438	2306	2177	2052	1931	1815	1703	1596	1494
2	2700	2681	2652	2613	2565	2510	2450	2384	2314	2240
3	1890	1964	2083	2090	2136	2176	2205	2225	2234	2240
4	0992	1087	1169	1254	1336	1414	1488	1557	1622	1680
5	0417	0476	0538	0602	0668	0735	0804	0872	0941	1008
6	0146	0175	0206	0241	0278	0319	0362	0407	0455	0504
7	0004	0055	0068	0083	0099	0118	0140	0163	0188	0216
8	0012	0015	0020	0025	0031	0039	0047	0057	0068	0081
9	0003	0004	0005	0007	0009	0011	0014	018	0022	0027
10	0001	0001	0001	0002	0002	0003	0004	0005	0006	0008

Продовження додатку В

m	*a*									
	3.5	**4.0**	**4.5**	**5.0**	**6.0**	**7.0**	**8.0**	**9.0**	**10.0**	**11.0**
0	0.0302	0183	0111	0067	0025	0009	0003	0001	-	-
1	1057	0733	0500	0333	0149	0064	0027	0011	0005	0002
2	1850	1465	1125	0842	0446	0223	0107	0050	0023	0010
3	2158	1954	1687	1404	0892	0521	0286	0150	0076	0037
4	1888	1954	1898	1755	1339	0912	0573	0338	0189	0102
5	1327	1563	1708	1755	1606	1277	0916	1277	0916	0224
6	0771	1042	1281	1462	1606	1490	1221	0911	0631	0411
7	0386	0595	0824	1044	1377	1490	1396	1171	0901	0646
8	0169	0298	0463	0653	1033	1304	1396	1318	1126	0888
9	0066	0132	0232	0363	0638	1014	1241	1318	1251	1058
10	0023	0053	0104	0181	0413	0710	0993	1186	1251	1194
11	0007	0019	0043	0082	0225	0452	0722	0970	1137	1194
12	0002	0006	0016	0034	0113	0264	0481	0728	0948	1094

Додаток Г

Випадкові величини

3393	6270	4228	6069	9407	1865	8549	3217	2351	8410
9108	2330	2157	7416	0398	6173	1703	8132	9065	6717
7891	3590	2502	5945	3402	0491	4328	2365	6175	7695
9085	6307	6910	9174	1753	1797	9229	3422	9861	8357
2638	2908	6368	0398	5495	3283	0031	5955	6544	3883
1313	8338	0623	8600	4950	5414	7131	0134	7241	0651
3897	4202	3814	3505	1599	1649	2784	1994	5775	1406
4380	9543	1646	2850	8415	9120	8062	2421	6161	4634
1618	6309	7909	0874	0401	4301	4517	9197	3350	0434
4858	4676	7363	8141	6133	0549	1972	3461	7116	1496
5354	9142	0847	5393	5416	6505	7156	5634	9703	6221
0905	6986	9396	3975	9255	0537	2479	4589	0562	5345
1420	0470	8679	2328	3939	1292	0406	5428	3789	2882
3218	9080	6604	1813	8209	7039	2086	3369	4437	3798
9697	8431	4387	0622	6893	8788	2320	9358	5904	9539
0912	4964	0502	9683	4636	2861	2876	1273	7870	2030
4636	7072	4868	0601	3894	7182	8417	2367	7032	1003
2515	4734	9878	6761	5636	2949	3979	8650	3430	0635
5964	0412	5012	2369	6461	0678	3693	2928	3740	8047
7848	1523	7904	1521	1455	7089	8094	9872	0898	7174
5192	2571	3643	0707	3434	6818	5729	8615	4298	4129
8438	8325	9886	1805	0226	2310	3675	5058	2515	2388
8166	6349	0319	5436	6838	2460	6433	0644	7428	8556
9158	8263	6504	2562	1160	1526	1816	9690	1215	9590
6061	3525	4048	0382	4224	7148	8259	6526	5340	4064

Додаток Д

Значення інтеграла ймовірності $\Phi(t) = \frac{2}{\sqrt{2\pi}} \int_0^t e^{-\frac{t^2}{2}} dt$

t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$
0.00	0.00000	0.30	0.23582	0.60	0.45149	0.90	0.63188
01	00798	31	24344	61	45814	91	63718
02	01596	32	25103	62	46474	92	64243
03	02393	33	25860	63	47131	93	64763
04	03191	34	26614	64	47783	94	65278
0.05	0.03988	0.35	0.27366	0.65	0.48431	0.95	0.65789
06	04784	36	28115	66	49075	96	66294
07	05581	37	28862	67	49714	97	66795
08	06376	38	29605	68	50350	98	67291
09	07171	39	30346	69	50981	99	67783
0.10	0.07966	0.40	0.31084	0.70	0.51607	1.00	0.68269
11	08759	41	31819	71	52230	01	68750
12	09552	42	32552	72	52848	02	69227
13	10348	43	33280	73	53461	03	69699
14	11134	44	34006	74	54070	04	70166
15	11924	45	34729	75	54675	1.05	70628
16	12712	46	35448	76	55275	06	71086
17	13499	47	36164	77	55870	07	71538
18	14285	48	36877	78	56461	08	71986
19	15069	49	37587	79	57047	09	72429
0.20	0.15852	0.50	0.38292	0.80	0.57629	1.10	0.72867
21	16633	51	38995	81	58206	11	73300
22	17413	52	39694	82	58778	12	73729
23	18191	53	40389	83	59346	13	74152
24	18967	54	41080	84	59909	14	74571
25	19741	55	41768	85	60468	15	74986
26	20514	56	42452	86	61021	16	75395
27	21284	57	43132	87	61570	17	75800
28	22052	58	43809	88	62114	18	76200
29	22818	59	44481	89	62653	19	76595

Продовження додатку Д

t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$
1.20	0.76986	1.55	0.87886	1.90	0.94257	2.25	0.97555
21	77372	56	88124	91	94387	26	97618
22	77754	57	88358	92	94514	27	97679
23	78130	58	88589	93	94639	28	97739
24	78502	59	88817	94	94762	29	97798
25	78870	1.60	0.89040	95	94882	2.30	0.97855
26	79233	61	89260	96	95000	31	97911
27	79592	62	89477	97	95116	32	97966
28	79945	63	89690	98	95230	33	98019
29	80295	64	89899	99	95341	34	98072
1.30	0.80640	65	90106	2.00	0.95450	35	98123
31	80980	66	90309	01	95557	36	98172
32	81316	67	90508	02	95662	37	98221
33	81648	68	90704	03	95764	38	98269
34	81975	69	90897	04	95865	39	98315
35	82298	1.70	0.91087	05	95964	2.40	0.98360
36	82617	71	91273	06	96060	41	98405
37	82931	72	91457	07	96155	42	98448
38	83241	73	91637	08	96247	43	98490
39	83547	74	91814	09	96338	44	98531
1.40	0.83849	75	91988	2.10	0.96427	45	98571
41	84146	76	92159	11	96514	46	98611
42	84439	77	92327	12	96599	47	98649
43	84728	78	92492	13	96683	48	98686
44	85013	79	92655	14	96765	49	98723
45	85294	1.80	0.92814	15	96844	2.50	0.98758
46	85571	81	92970	16	96923	51	98793
47	85844	82	93124	17	96999	52	98826
48	86113	83	93275	18	97074	53	98859
49	86378	84	93423	19	97148	54	98891
1.50	0.86639	1.85	0.93569	2.20	0.97219	2.55	0.98923
51	86696	86	93711	21	97289	56	98953
52	87149	87	93852	22	97358	57	98983
53	87398	88	93989	23	97425	58	99012
54	87644	89	94124	24	97491	59	99040

Продовження додатку Д

t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$
2.60	0.99068	2.95	0.99682	3.30	0.99903	3.65	0.99974
61	99095	96	99692	31	99907	66	99975
62	99121	97	99702	32	99910	67	99976
63	99146	98	99712	33	99913	68	99977
64	99171	99	99721	34	99916	69	99978
65	99195	3.00	0.99730	35	99919	3.70	0.99978
66	99219	01	99739	36	99922	71	99979
67	0.99241	02	99747	37	99925	72	99980
68	99263	03	99755	38	99928	73	99981
69	99285	04	99763	39	99930	74	99982
2.70	0.99307	05	99771	3.40	0.99933	75	99982
71	99327	06	99779	41	99935	76	99983
72	99347	07	99786	42	99937	77	99984
73	99367	08	99793	43	99940	78	99984
74	99386	09	99800	44	99942	79	99985
75	99404	3.10	0.99806	45	99944	3.80	0.99986
76	99422	11	99813	46	99946	81	99986
77	99439	12	99819	47	99948	82	99987
78	99456	13	99825	48	99950	83	99987
79	99473	14	99831	49	99952	84	99988
2.80	0.99489	15	99837	3.50	99953	85	99988
81	99505	16	99842	51	99955	86	99989
82	99520	17	99848	52	99957	87	99989
83	99535	18	99853	53	99958	88	99990
84	99549	19	99858	54	99960	89	99990
85	99563	3.20	0.99863	55	99961	3.90	0.99990
86	99576	21	99867	56	99963	91	99991
87	99590	22	99872	57	99964	92	99991
88	99602	23	99876	58	99966	93	99992
89	99615	24	99880	59	99967	94	99992
2.90	0.99627	25	99855	3.60	0.99968	95	99992
91	99639	26	99889	61	99969	96	99992
92	99650	27	99892	62	99971	97	99993
93	99661	28	99896	63	99972	98	99993
94	99672	29	99900	64	99973	99	99993

Навчальне електронне видання

БУРГАЗ Марина Іванівна

СТАТИСТИЧНІ МЕТОДИ В БІОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

Конспект лекцій

Видавець і виготовлювач
Одеський державний екологічний університет
вул.Львівська, 15, м. Одеса, 65016
тел./факс; (0482) 32-67-35 E-mail: info@odeku.edu.ua
Свідоцтво суб'єкта видавничої справи
ДК № 5242 від 08.11.2016