МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ДЕРЖАВНИЙ ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД
«ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ»

КОНСПЕКТ ЛЕКЦІЙ

з навчальної дисципліни циклу за вибором ВНЗ

БІОМЕТРІЯ

напрям підготовки 6. 040106 «Екологія, охорона навколишнього середовища та збалансоване природокористування»

ЗМІСТ

Біометрія - наука про статистичний аналіз групових властивостей біологічних і екологічних об'єктів. Під статистичним аналізом мається на увазі сукупність постулатів і методів теорії ймовірності та математичної статистики застосовуваної в даному випадку до особливостей біологічних і екологічних об'єктів.

У практиці біометричних досліджень використовується своя специфічна термінологія. Деякі з цих термінів та їх відповідностей математичним наведені в табл. 1.1.

Таблиця 1.1 – Математичні та біометричні терміни

Математика	Біометрія
1. Величина	1. Дата, признак
2. Середнє значення	2. Середня величина признаку
3.Сума квадратів центральних відхилень	3. Дисперсія
4. Середній квадрат	4. Варіанса
5. Розсіювання, варіабельність, розкид	5. Різноманітність, мінливість

Групові властивості діляться на дві категорії - основні й сполучені.

До основних групових властивостей відносяться середній рівень ознаки, який характерний для всієї групи в цілому.

Парні групові властивості - такі групові властивості, які з'являються внаслідок розвитку основних властивостей.

Різноманітність ознаки - неминуча неоднаковість, більша або менша відмінність особин в групі по досліджуваному ознакою.

1. СЕРЕДНІ ВЕЛИЧИНИ ТА ЇХ ВИДИ

Основним показником групових властивостей в біометрії є середня величина, яка широко використовується в науці і практиці. При вивченні рослин, тварин, мікроорганізмів і людини розрахунок середніх показників становить основу обробки первинного матеріалу.

Середні розміри особин чисельність і їх маса служать для характеристики видів, різновидів, сортів, порід та інших біологічних груп. Середні показники фізіологічних процесів характеризують інтенсивність різних сторін внутрішнього обміну організмів або силу дії біологічних агентів і медичних препаратів.

У виробництві середні показники стали звичайними характеристиками оцінки роботи окремих фахівців, господарств, областей.

Середня величина якої-небудь ознаки визначається для того, щоб отримати характеристику цієї ознаки для всієї досліджуваної групи в цілому.

Середня величина ознаки визначається різними способами залежно від об'єкта спостереження і поставлених цілей. Тому є не один, а кілька видів середніх, що призводить до визначення типу середньої:

арифметичної;

геометричній;

гармонійної;

квадратической і т.д.,

а також до загальноприйнятого розмежування між аналітичними і позиційними середніми .

До аналітичних середнім відносяться всі середні величини, які виражаються за допомогою формул.

Позиційні - всі середні (медіана, розділову значення, мода, вище і нижче значення), які визначаються по відношенню до загального розподілу величин.

Для біологів і екологів найбільше значення мають чотири аналітичних середніх: середня арифметична, середня геометрична, середня квадратична і середня гармонійна. Крім того, для характеристики біологічних сукупностей вживаються неаналітичні позиційні середні: мода, медіана, розділову значення, ліміти, квартили, децили і т.д.

1.1 Класифікація середніх величин

У загальному випадку середня величина може бути віднесена до наступних категорій:

рухома або стійка;

базова або експонентна або базово-експонентна;

одноплощинна або двуплощинна;

однозначна або багатозначна;

з повною або неповною областю застосування.

Рухомою називається середня, яка залежить від величини всіх членів вибірки, так що із зміною будь-якого з них змінюється і величина середньої.

На противагу рухомій - стійка середня є полусумою крайніх членів. Якщо, наприклад, середньоарифметична може змінюватися, то полусума крайніх членів залишається незмінною.

Аналітичні середні діляться на базові, експонентні і базовоекспоненціальні залежно від того чи фігурують їхні члени вибірки в математичних формулах, які їх висловлюють, в якості підстави та/або показника ступеня. Наприклад, базовими середніми є арифметична і геометрична.

Базові середні можуть бути одноплощіними або двуплощінимі, залежно від того, чи входять члени сукупності тільки в чисельник або знаменник або одночасно і в чисельник, і в знаменник.

Середня називається однозначною, якщо, які б не були члени вибірки за їх кількістю і величиною (позитивні або негативні), середня розглянутого типу має лише одне значення.

Багатозначні середні мають кілька значень, наприклад середньогеометрична.

Якщо середня може бути визначена, яка б не була сукупність розглянутих ознак, то вважається, що область застосування середньої повна, інакше говорять про неповною області застосування.

1.2 Загальна формула середніх величин

Чотири основні види середніх величин можна виразити єдиною формулою:

$$Ср_в = \sqrt[m]{\frac{\sum_{i=1}^{n} P_i \cdot V_i^m}{\sum_{1=1}^{n} P_i}}$$

В окремому випадку, коли n $P_1 = P_2 = \ldots = P_n$

$$Ср = \sqrt[m]{\frac{\sum_{i=1}^{n} V_i^m}{n}}$$

де Ср, $Ср_В$ - середня величина проста та зважена; V - дата (ознака), окреме значення досліджуваної ознаки у кожного об'єкта дослідження; m - показник, що визначає

вид середньої; n - число дат, що усереднюється (ознак); Pi - математичний вага (значимість) ознаки у вибірці.

Надаючи показником m різні значення, наприклад: 1, 2, -1, 0, можна отримати формули для окремих видів середніх.

При m = 1 отримуємо формулу середньої арифметичної:

$$M = \frac{\sum_{i=1}^{n} V_i}{n}$$

При m = 2 отримуємо формулу середньої квадратичної

$$S = \sqrt{\frac{\sum_{i=1}^{n} V^2}{n}}$$

При m = -1 отримуємо формулу середньої гармонійної:

$$H = \frac{n}{\sum_{i=1}^{n} V_i^{-1}}$$

При m = 0, після спеціальних перетворень, отримуємо формулу середньої геометричної:

$$G = \sqrt[n]{\prod_{i=1}^{n} V_i}$$

Якщо в загальну формулу середньої підставити m = - ∞ і m = + ∞, то після перетворень отримаємо два крайніх значення в групі: min - найменше значення і max - найбільше значення.

1.3 Загальні властивості середніх величин

Для правильного застосування середніх величин необхідно знати наступні властивості цих показників: серединна розташування, абстрактність і єдність сумарного дії.

За своїм чисельним значенням всі середні величини займають проміжні положення між мінімальним і максимальним значеннями ознаки.

При цьому найменшу величину має середня гармонійна, а найбільшу - середня квадратична. Наступна схема показує положення кожної середньої по відношенню один до одного:

m = { -∞ -1 0 +1 +2 +∞ }

Ср = { min< H < G < M < S < max }

Облік зазначених взаємовідносин між середніми величинами допомагає при перевірці вироблених обчислень. Наприклад, якщо середня арифметична виявилася вище максимального значення ознаки або якщо середня геометрична більше середньої арифметичної, то, очевидно, що в розрахунках є помилки.

Серединне розташування. Середня ознаки показує, яку величину мав би кожен з представників досліджуваної групи, якби всі вони були однаковими і їх сумарна дія була такою ж, як і від фактичних не усереднених значень цієї групи. При використанні середніх величин передбачається, що поки вони застосовуються, різнорідна група замінена однорідною групою, в якій всі значення ознаки однакові і рівні середній величині.

Наприклад, якщо є п'ять значень ознаки: 1; 4; 5; 5; 5 з середньою величиною M = 4, то при використанні цієї середньої передбачається, що різнорідна група замінена на однорідну з однаковими значеннями: 4; 4; 4; 4; 4.

Дана особливість середніх величин лежить в основі таких звичайних виробничих виразів як «від кожної корови одержано по 3000 л молока» , «з кожного гектара зібрано по 500 ц буряків», «з кожного вулика отримано по 80 кг меду», «при відгодівлі отримано по 100 кг приросту на кожну голову» і т.п. Корови дають, звичайно, різні удої, на різних ділянках отримано різний врожай і т.д., але все ж для виробничої характеристики господарства і, особливо, для планових розрахунків виявилося зручним умовно прийняти, що всі корови дали або даватимуть однаковий удій, рівний середній величині цієї ознаки для даного стада та року (« від кожної корови»), або, що з кожного гектара отримано один і той же урожай, рівний середньому врожаю з загальної площі (« з кожного гектара»).

Абстрактність. Замінити різнорідну групу однорідною можна тільки шляхом відволікання від тих відмінностей, які існують в дійсності. Тільки абстрагуючись від наявних індивідуальних різноманітних значень, можна дати необхідну характеристику групи одним числом - середньою величиною ознаки. У цьому сенсі всяка середня величина є насамперед абстрактна величина, яка часто насправді не існує, а іноді і не може існувати.

Єдність сумарного дії. Не всяке вирівнювання відмінностей у групі може привести до правильної середній величині. Обчислення середніх величин необхідно вести таким чином, щоб сумарна дія вирівняних значень ознаки дорівнювала б сумарній дії первинних не усереднених значень.

Наприклад, якщо чотири дорослих особини якої-небудь промисловий птиці важили 2; 3; 3; 4 кг, то середня вага цих птахів:

$$(2+3+3+4)/4 = 3 \text{ кг}.$$

Сумарна вага чотирьох усереднених значень 3 + 3 + 3 + 3 = 12 кг. Така ж сумарна вага малася і в дійсності: 2+3+3+4=12 кг. В даному випадку вибір як середньої - середньої арифметичної зроблений правильно, але так буває не завжди.

Наприклад, потрібно розрахувати середньорічний приріст популяції якогось виду за два роки, якщо відомо, що за перший рік приріст склав 20 %, а за другий - 60 % (від початку другого року). Використовуючи метод середньої арифметичної, отримуємо:

$$M = \frac{20\% + 60\%}{2} = 40\%$$

В даному випадку, застосування цієї середнього не буде правильним, так як два усереднених значення у своєму сумарному дії не дадуть того ж результату, який дали два фактичних не усереднених значення. Фактичний загальний сумарний приріст популяції за два роки визначається наступним чином.

До кінця першого року популяція становить:

$$100\% + \frac{100 \cdot 20\%}{100\%} = 120\%$$

$$120\% + \frac{120\% \cdot 60\%}{100\%} = 192\%$$

а приріст від початку 1-го і до кінця 2-го року складе 192% - 100% = 92%

Якщо ж взяти за середній приріст розраховану раніше величину 40 %, то до кінця 1-го року популяція складе:

$$100\%+\frac{100\%\cdot 40\%}{100\%}=140\%$$

а до кінця 2-го року:

$$140\%+\frac{140\%\cdot 40\%}{100\%}=196\%$$

а приріст за два роки складе 196 % - 100 % = 96 %

Якщо ж використовувати середню геометричну, то середній приріст визначиться наступним чином:

$$G = \sqrt[2]{120\cdot 160} -100= 38,6\ \%$$

Чисельність популяції за 2 роки:

$$100\% +\frac{100\%\cdot 38.6\%}{100\%}=138.6\%$$

$$138,6\% +\frac{138.6\%\cdot 38.6\%}{100\%}=192\%$$

Таким чином, сумарний результат буде дорівнює фактичному: 192 % -100 % = 92%.

1.4. Аналітичні середні величини

1.4.1 Середня арифметична

Найпоширенішим показником середньої якості є середня арифметична. Обчислюється вона, за формулою:

$$M=\frac{\sum_{i=1}^{n}V_i}{n}$$

де M - середня арифметична, V - дата, окреме значення досліджуваного ознаки, n - число використаних значень ознаки. У розгорнутому вигляді ця формула має наступний вигляд:

$$M=\frac{V_1+V_2+V_3+\ldots+V_n}{n}$$

Часто застосовується при усередненні паралельних спостережень за величиною ознаки, наприклад, як маса, довжина, висота тіла, вміст речовини.

Приклад. Три паралельних визначення вмісту діоксиду сірки у трьох цехах заводу проводилося трьома різними лаборантами. Один виміряв = 0,75. Інший = 0,80. Третій = 0,70.

Середньоарифметична $M=\frac{0,75+0,80+0,70}{3}=0,75$

1.4.2 Зважена середня арифметична

Щоб розрахувати середню арифметичну, складають всі значення ознаки і отриману суму ділять на число дат. У цьому випадку кожне значення входить в суму однаковим чином, збільшуючи її на повну свою величину. Але таке не завжди є коректним. Іноді значення ознаки повинні входити в суму з неоднаковою (індивідуальної) поправкою.

Ця поправка, виражена певним множником, називається математичним вагою значення.

Середня, розрахована для значень ознаки з неоднаковими вагами, називається зваженою середньою. Зважена середня арифметична розраховується за формулою:

$$M_{взв} = \frac{\sum_{i=1}^{n}(V_i \cdot P_i)}{\sum_{i=1}^{n} P_i}$$

де V - значення ознаки, дата; P - математична вага або значимість значення, що усереднюється. Часто P є кількістю значень з даною величиною ознаки.

Щоб розрахувати зважену середню арифметичну, необхідно кожне значення ознаки помножити на його вагу, всі ці добутки скласти і отриману суму розділити на суму ваг.

Приклад. Є результати двох досліджень ваги бджіл: в одному випадку отримана середня величина 660 мг, в іншому - 600 мг. Потрібно отримати загальну середню вагу, причому відомо, що в першому дослідженні було виміряна вага у 100 бджіл, а в другому - у 20.

У даному випадку значеннями ознаки є середні 660 і 600 мг; їх вагами - чисельності груп $p_1 = 100$ і $p_2 = 20$. Зважена середня арифметична розраховується наступним чином:

$$M_{зв} = \frac{660 \cdot 100 + 600 \cdot 20}{100 + 20} = 650 \text{ мг}$$

У разі простою середньоарифметичної було б отримано:

$$M = \frac{660 + 600}{2} = 630 \text{ мг},$$

що є не коректним значенням.

1.4.3 Середня геометрична

Щоб отримати середню геометричну для групи з n датами потрібно всі дати перемножити і з отриманого добутку витягти корінь n-го ступеня:

$$G = \sqrt[n]{\prod_{i=1}^{n} V_i}$$

або через логарифмічну форму:

$$G = \exp\left(\frac{1}{n} \cdot \sum_{i=1}^{n} \ln V_i\right)$$

де G - середня геометрична; n - число дат; П - добуток дат V_i; ln - натуральний логарифм для кожної з дат V_i.

Приклад. Є не усереднена група ознак: 1; 4; 5; 5; 5. Знайти середньогеометричну цих величин.

$$G = \sqrt[5]{1 \cdot 5 \cdot 5 \cdot 5 \cdot 4} = 3{,}4654$$

$$\prod = 1 \cdot 4 \cdot 5 \cdot 5 \cdot 5 = 500$$

Якщо деякі з членів сукупності позитивні, а інші негативні, то можна отримати одну - дві геометричні середні, а також може не бути жодної. Якщо є два значення, то зазвичай приймають як среднегеометричну ту з них, знак якої збігається зі знаком середньої арифметичної.

При підрахунку среднегеометричної не повинно бути нульових значень дат.

Застосовується середня геометрична у всіх випадках, коли необхідно дізнатися чи спланувати середні прирости за певний період. При розрахунках середнього поперіодного приросту можливі два основних способи застосування середньої геометричної.

Перший спосіб застосовується, коли є відомості про приростах за кожен період, виражених у відсотках або частках (відсоток, поділений на 100) від початку кожного періоду. У таких випадках розрахунок середнього приросту ведеться за формулою:

$$x = \sqrt[n]{\prod_{i=1}^{n}(1+a_i)} - 1$$

або для дат виражених у відсотках:

$$x = \sqrt[n]{\prod_{i=1}^{n}(100+a_i)} - 100,\ \%$$

де x - середній поперіодний приріст за ряд періодів рівної тривалості, a - фактичний приріст за той чи інший період, виражений в частках або відсотках, n - число періодів.

З цієї формули випливає, що для знаходження середнього приросту за першим способом потрібно частку фактичного приросту за кожний період додати до одиниці, отримані величини перемножити і з добутку витягти корінь n-го ступеня, а потім відняти одиницю.

Приклад: Поголів'я кроликів в господарстві збільшилася за 1 -й рік на 5 %, за 2 -й - на 20 %, за 3 -й - на 50 %, за 4 -й - на 50 %, вважаючи кожного разу від минув року.

Розрахувати середньорічний приріст популяції за ці роки.

Визначимо через середньогеометричну:

$$x = \sqrt[4]{105 \cdot 120 \cdot 150 \cdot 150} - 100 = 29{,}76\%$$

Загальний приріст за минулий період років можна визначити за такою формулою:

$$G_{\Sigma} = \left(\left(1 + \frac{x}{100}\right)^{n} - 1\right) \cdot 100$$

Наприклад, для попереднього прикладу:

$$G_{\Sigma}=\left(\left(1+\frac{29.76}{100}\right)^{4}-1\right)\cdot 100=377,4\%$$

Другий спосіб розрахунку середніх приростів застосовується в тих випадках, коли є дані про абсолютні кількості особин (ознак, об'єктів) на початок і кінець загального великого періоду і потрібно розрахувати середній приріст за більш дрібні періоди. У таких випадках середній приріст розраховується за формулою:

$$x=\sqrt[n]{\frac{A_n}{A_1}}-1 \text{ чи } x=\left(\sqrt[n]{\frac{A_n}{A_1}}-1\right)\cdot 100\%$$

де x - середній приріст за більш дрібні періоди, A_n - кількість особин на кінець загального періоду, або на кінець останнього n-го дрібного періоду, A_1 - кількість особин на початок досліджуваного загального періоду, або на початок 1-го дрібного періоду.

Величину ознаки A_n (наприклад, кількість особин) на кінець загального періоду, або на кінець останнього n-го дрібного періоду років можна визначити за такою формулою:

$$A_n=A_1\cdot\left(1+\frac{x}{100}\right)^{n}$$

Приклад. В агропромисловому господарстві на початок п'ятирічного періоду було 100 вуликів, а до кінця стало 140. Визначити середньорічний відсоток збільшення пасіки за минулі п'ять років.

Застосовуючи зазначену вище формулу, отримаємо:

$$x=\sqrt[5]{\frac{140}{100}}-1=0,0697 \text{ чи } 6,97\%$$

Приклад. В області заплановано за п'ять років збільшити обсяги переробки твердих побутових відходів (ТПВ) на 60 %. Потрібно розподілити це завдання рівномірно по роках.

У даному випадку не задані абсолютні кількості на початку і кінці загального періоду, але даний загальний відсоток приросту за весь період - 60 %, що дає можливість легко отримати необхідне ставлення A_n/A_1. Обсяг переробки ТПВ повинен збільшитися на 60 %. Це означає, що на кожні 100 одиниць, що були на початку загального періоду, повинно бути 160 одиниць в кінці. Для виконання такого завдання середньорічний приріст можна запланувати такий спосіб:

$$x=\sqrt[5]{\frac{160}{100}}-1=0,0985$$

Виявилося, що для збільшення переробки ТПВ за п'ятирічку на 60 % достатньо забезпечити середньорічний приріст на 9,85 %, а не ~ 12%:

$$\frac{60\%}{5}=12\%$$

як це могло здатися без урахування того, що середній приріст утворюється за принципом середньої геометричної, а не середньої арифметичної.

1.4.4 Зважена середня геометрична

Визначається за формулою:

$$G_{взв}=\sqrt[\sum_{i=1}^{n}P_i]{\prod_{i=1}^{n}(V_i+100)^{P_i}}-100,$$

або через логарифмічну форму:

$$G_{взв}=\exp\left(\frac{1}{\sum_{i=1}^{n}P_i}\cdot\sum_{i=1}^{n}(P_i\cdot\ln V_i)\right)-100,$$

де Vi - величина ознаки виражена в %.

Приклад. Вибірка приросту ваги різних груп телят через один місяць приведена в табл.1.2. Визначити середній приріст телят, використовуючи зважену середньогеометричну:

Таблиця 1.2

V, %	10	14	17	21
P, шт	3	11	9	2

$$\sum_{i=1}^{n}P_i=3+11+9+2=25$$

$$G_{взв}=\sqrt[25]{(10+100)^3\cdot(14+100)^{11}\cdot(17+100)^9\cdot(21+100)^2}-100=15{,}13\text{ кг}$$

1.4.5 Середня і зважена середня квадратична

Середня квадратична обчислюється за формулою:

$$S=\sqrt{\frac{\sum_{i=1}^{n}V^2}{n}}$$

тобто вона дорівнює кореню квадратному із суми квадратів дат, поділеній на їх число.

Приклад. Якщо є п'ять дат: 1, 4, 5; 5; 5, то середня квадратична буде:

$$S=\sqrt{\frac{1^2+4^2+5^2+5^2+5^2}{5}}=4{,}3$$

На відміну від інших середніх, тут виходять найбільш завищені значення. Вживається середня квадратична, наприклад, при розрахунку середніх радіусів (діаметрів) кіл з метою визначення середньої площі чого-небудь.

Приклад. Вимірювання діаметрів колоній, отриманих від посіву мікробів певного виду, дали такі результати (у мм): 15; 20; 10; 25; 30. Визначити середню площу посівів колоній.

Використовуємо середню квадратичну:

$$S = \sqrt{\frac{15^2 + 20^2 + 10^2 + 25^2 + 30^2}{5}} = 21{,}22 \text{ мм}$$

Середня площа буде дорівнює:

$$C_{ср} = \frac{\pi \cdot D^2}{4} = \frac{\pi \cdot 21{,}22^2}{5} = 353{,}7 \text{ мм}^2$$

Порівняємо із середньоарифметичним значенням:

$$M = \frac{15 + 20 + 10 + 25 + 30}{5} = 20 \text{ мм}$$

Використовую принцип єдності сумарного дії перевіримо яка з отриманих середніх величин є коректною. Для цього розрахуємо загальну площу всіх колоній.

Загальна площа всіх колоній через середню арифметичну складе:

$$C_s = 5 \cdot \frac{\pi \cdot M^2}{4} = 5 \cdot \pi \frac{20^2}{4} = 1570 \text{ мм}^2$$

Насправді загальна площа буде дорівнює:

$$C_s = \frac{\pi}{4} \cdot \sum_{i=1}^{n} D^2 = n \cdot \frac{\pi}{4} \cdot S^2 = 5 \cdot \frac{\pi}{4} \cdot 21{,}22^2 = 1767{,}4 \text{ мм}$$

Таким чином, коректною середньою є середня квадратична.

Зважена середня квадратична визначається за формулою:

$$S = \sqrt{\frac{\sum_{i=1}^{n} p_i \cdot V_i^2}{\sum_{i=1}^{n} p_i}}$$

1.4.6 Середня гармонійна

Середня гармонійна розраховується за формулою:

$$H = \frac{n}{\sum_{i=1}^{n} V_i^{-1}}$$

Наприклад, для п'яти дат 1; 4; 5; 5; 5 її можна визначити наступним чином:

$$H = \frac{5}{\frac{1}{1}+\frac{1}{4}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5}} = 2{,}7$$

Застосовується середня гармонійна при усередненні ознак, за допомогою яких надалі знаходять величини (інші ознаки) обернено пропорційні (або питомі) по відношенню до первинних ознаками. Наприклад, мінливих швидкостей, з метою подальшого розрахунку середнього або загального часу руху або протікання процесу; середнього обсягу з метою подальшого розрахунку середньої щільності і т.д.

У цьому випадку, до дат пред'являються такі основні вимоги:

- при застосуванні середньогармонічної не повинно бути нульових дат;
- якщо є й негативні і позитивні значення, необхідно щоб знаменник не дорівнював нулю.

Приклад. Поштові голуби однієї станції до місця годування летять зі швидкістю 50 км/год, а у зворотному напрямку - зі швидкістю 40 км/год. Якщо крім цих даних, нічого більше невідомо і потрібно з'ясувати середню швидкість польоту для обох напрямів (відстані рівні), то зробити це можна, розрахувавши просту середню гармонійну для двох дат - 50 і 40:

$$H = \frac{2}{\frac{1}{40}+\frac{1}{50}} = 44{,}4 \text{ км/год}$$

1.4.7 Зважена середня гармонійна

Визначається за формулою:

$$H_{зв} = \frac{\sum_{i=1}^{n} P_i}{\sum_{i=1}^{n} \frac{P_i}{V_i}},$$

де суму ваг ділять на суму співвідношень ваг з відповідними датами.

Зважування швидкостей в цьому випадку проводиться по чисельнику питомих величин ознак, наприклад, для усереднює швидкостей - по відстанях.

Приклад. Рисак на тренуваннях пробігав одну за одною три дистанції, різні за станом дороги. Швидкість на першій дистанції складала 13 км/год, на другому - 20 км/год і на третьому - 10 км/год. Відомо, що перша дистанція була в 2 рази, а друга в 4 рази довше третьої. За цими даними знайти середню швидкість рисака.

$$H_{зв} = \frac{2+4+1}{\frac{2}{13}+\frac{4}{20}+\frac{1}{10}} = 15{,}4 \text{ км/год}$$

1.5 Прості неаналітичних (позиційні) середні

Неаналітичних (позиційні) середні - це такі величини розглянутої ознаки, місце розташування яких в даній вибірці не може бути виражене у вигляді аналітичної функції або у вигляді сукупності алгебраїчних членів, але, які знаходяться в залежності від положення членів впорядкованої послідовності вибірки, яку можна утворити. З цих членів і можуть бути визначені середні положення на основі того, як вони розташовані по відношенню до деяких або до всіх членів сукупності.

У практиці еколого-біологічних досліджень найбільше застосування отримали наступні позиційні середні:

медіана;
квартили;
децили;
перцентілі (центів);
квантилі;
розділову значення;
мода.

1.5.1 Медіана

Медіана - таке значення ознаки, яке поділяє монотонно зростаючу групу на дві рівні частини. Одна частина містить значення ознаки менше ніж медіана, а інша - більші значення.

Приклад: 1 2 3 4 5 6 7 8 9

Медіана = 5. При цьому число членів праворуч і ліворуч від медіани дорівнюють по 4.

Коли число ознак непарне вибирається центральне значення ознаки, яке ділить групу навпіл.

Якщо число членів парне, то будь-яка величина, що знаходиться між значеннями двох центральних членів ділить послідовність навпіл і може бути прийнята за медіану. Зазвичай за медіану беруть середню арифметичну двох центральних членів.

Якщо ми додамо до розглянутої послідовності ще одну дату, то отримаємо:

1 2 3 4 5 6 7 8 9 10

$$\text{Медіана} = \frac{5+6}{2} = 5{,}5$$

У загальному вигляді, коли число членів ряду парне, медіану X визначають з рішення наступного рівняння:

$$(X-V_1)\,(X-V_2)\cdot\ldots\cdot(X-V_i)=(V_{i+1}-X)\,(V_{i+2}-X)\ldots\cdot(V_n-X)$$

Якщо n = 2, то:

$$X=\frac{V_1+V_2}{2}$$

Якщо n = 4, то:

$$X=\frac{V_4\cdot V_3-V_2\cdot V_1}{(V_4+V_3)-(V_1+V_2)}$$

Для численних груп медіану можна розрахувати за формулою:

$$M_e = W_H + k\cdot\left(\frac{\frac{n}{2}-\sum}{f}\right),$$

де M_e - медіана; Wн - початок класу, в якому знаходиться медіана; k - величина класового проміжку; n - загальне число дат в групі; Σ - сума частот класів (починаючи з меншого), що передують класу, в якому знаходиться медіана; f - частота класу, в якому знаходиться медіана.

1.5.2 Квартилі [Qi (i = 1, 2, 3)]

Якщо маємо монотонний зростаючий ряд ознак:

$V_1, V_2, V_3, \ldots, V_n$

і два цілих позитивних числа p і q, сума яких дорівнює 4, то квартилями називаються конкретні величини, що задовольняють умові, що число членів ряду, попередніх їм, дорівнює $\left(\frac{p}{q}\right)$ числу членів, наступних за ними. Таким чином, будемо мати:

1-я квартиль Q_1, якщо $\frac{p}{q} = \frac{1}{3}$, тобто ділить вибірку у відношенні 1 до 3;

2-я квартиль Q_2, якщо $\frac{p}{q} = \frac{2}{2}$, друга квартиль збігається з медіаною;

3-тя квартиль Q_3, якщо $\frac{p}{q} = \frac{3}{1}$, третя ділить вибірку у співвідношенні 3 до 1.

Таким чином, всі три квартилі розбивають вибірку на 4-ри рівні частини.

Квартилі визначаються таким чином:

- якщо вираз $t = \frac{i \cdot n}{4}$ (i= 1,2,3) не є цілим числом, то і-я квартиль є членом x_t ряду, де t - найменше з цілих чисел, що перевищують цей вираз (округлене до цілого в більшу сторону).

- якщо ж відношення $\frac{i \cdot n}{4}$ - ціле число, то і-я квартиль може бути представлена кожним числом, що містяться в інтервалі між x_t і x_{t+l}, і, зокрема, може бути дорівнює середньої арифметичної цих двох членів ряду.

Приклад. Якщо дано ряд ознак:

t =	1	2	3	4	5
V =	2	5	8	16	24

то при i = 1 (перша квартиль) величина $t = \frac{i \cdot n}{4} = \frac{1 \cdot 5}{4} = 1{,}25$ не є цілим числом. Саме менше з цілих чисел, що перевищують це відношення t = 1,25 ≈ 2. Отже, перша квартиль дорівнює другому члену ряду, тобто $Q_1 = x_2 = 5$.

Аналогічним чином знайдемо, що друга квартиль дорівнює:

$(t = \frac{i \cdot n}{4} = \frac{2 \cdot 5}{4} = 2{,}5 \approx 3)$, $Q_2 = x_3 = 8$,

а третя: ($t=\frac{i \cdot n}{4}=\frac{3 \cdot 5}{4}=3{,}75 \approx 4$), $Q_3=x_4=16$,

Приклад. Якщо дано ряд чисел:

t=	1	2	3	4	5	6	7	8
V=	2	5	8	16	24	28	30	32

то при i = 1, 2 і 3, ми отримаємо для відносини $t=\frac{i \cdot n}{4}$ наступні значення:

$$\frac{1 \cdot 8}{4}=2 \quad \frac{2 \cdot 8}{4}=4 \quad \frac{3 \cdot 8}{4}=6$$

Таким чином, всі три значення - цілі.

У цьому випадку перша квартиль буде знайдена як середня арифметична другого і третього членів ряду, друга квартиль - як середня арифметична четвертого і п'ятого членів і третя квартиль - як середня арифметична шостого та сьомого членів. Таким чином будемо мати:

$$Q_1=\frac{5+8}{2}=6{,}5 \quad Q_2=\frac{16+24}{2}=20 \quad Q_3=\frac{28+30}{2}=29$$

Слід зазначити, що в цьому випадку перша квартиль є медіаною членів, що передують медіані всього ряду, а третя квартиль є медіаною наступних за нею членів.

Приклад. Якщо дано ряд:

t=	1	2	3	4	5	6	7
V=	2	5	8	16	24	28	30

то для відносини $t=\frac{i \cdot n}{4}$ отримаємо такі значення:

$$\frac{7}{4}=1{,}75 \quad \frac{14}{4}=3{,}5 \quad \frac{21}{4}=5{,}25$$

Таким чином, всі три значення - не цілі. Найменші цілі числа, що перевершують відповідно ці три значення, будуть: 2, 4, 6. Отже, квартилі будуть рівні 2-му, 4-му і 6-му членам ряду, т. б. $Q_1 = 5$, $Q_2 = 16$, $Q_3 = 28$.

Однак цей останній випадок значно відрізняється від попередніх. У них число членів ряду, що передували першому квартилю, становило рівно одну третину числа членів, що слідували за нею, чого немає в даному випадку. У той же час ми не бачимо іншого способу точно визначити першу квартиль. Те ж саме можна сказати і про третю квартиль. Однак, хоча знайдені нами значення і не є тими величинами, які ми шукали, виходячи з даного нами визначення квартилей, практична необхідність змушує нас приймати їх як такі. Втім, ці розбіжності втрачають своє значення, коли число членів ряду велике.

Як і в другому прикладі, в даному випадку знайдена нами перша квартиль є медіаною членів, що передують медіані всього ряду, а третя квартиль - медианою наступних за нею членів.

1.5.3 Децилі [D_i (i = 1, 2, 3,...,9)]

Якщо дано монотонно зростаючий ряд ознак:

$$V_1, V_2, V_3, \ldots, V_n$$

і два цілих позитивних числа p і q, сума яких дорівнює 10, децилями називаються такі конкретні величини, які задовольняють умові, що число членів ряду, що передують децилям, дорівнює $\left(\frac{p}{q}\right)$ числу членів, наступних за ними. Іншими словами 9 децилів розбивають вибірку на 10 рівних частин таким чином:

1-я дециль D_1, якщо $\frac{p}{q} = \frac{1}{9}$;

2-я » D_2, » $\frac{p}{q} = \frac{2}{8}$;

3-я » D_3, » $\frac{p}{q} = \frac{3}{7}$;

4-я » D_4, » $\frac{p}{q} = \frac{4}{6}$;

5-я » D_5, » $\frac{p}{q} = \frac{5}{5}$;

6-я » D_6, » $\frac{p}{q} = \frac{6}{4}$;

7-я » D_7, » $\frac{p}{q} = \frac{7}{3}$;

8-я » D_8, » $\frac{p}{q} = \frac{8}{2}$;

9-я » D_9, » $\frac{p}{q} = \frac{9}{1}$;.

Всі децилі є членами ряду, якщо n - 1 кратно 10. Інакше - є середніми між прикордонними членами ряду. Визначаються децилі таким же методом, як і квартили.

1.5.4 Центилі [Ci (i = 1, 2, ..., 99)]

Якщо дано монотонно зростаючий ряд ознак:

$V_1, V_2, V_3, \ldots, V_n$

і два цілих позитивних числа p і q, причому p + q = 100, то центилями називаються величини, що задовольняють умові, що число попередніх їм членів ряду дорівнює $\left(\frac{p}{q}\right)$ числа членів, наступних за ними.

Для першої центилі C_1 маємо $\left(\frac{p}{q}\right)=\frac{1}{99}$, для другої центилі $\left(\frac{p}{q}\right)=\frac{1}{98}$ і т.д. Центилі знаходяться таким же методом, як квартилі і децилі.

1.5.5 Квантилі [Q_{ki} (i = 1, 2,..., k-1)]

Квантилі є узагальненням квартилей, децилів і центилей. Для двох цілих позитивних чисел p і q, причому p + q = k, квантилями називаються величини, що задовольняють умові, що число попередніх членів вибірки дорівнює $\left(\frac{p}{q}\right)$ числа наступних членів. При k = 4 ми маємо квартилі, при k = 10 - децилі і при k = 100 - центилі. Квантилі знаходяться таким же методом, як було розглянуто вище.

1.5.6 Роздільне значення (R_z)

Роздільним значенням вибірки у вигляді монотонно зростаючого ряду ознак називається таке середнє значення $R_z = V_k$, яке ділить його на дві частини, що задовольняють умові, що сума однієї частини з R_z повинна перевищувати суму іншій частині без R_z:

$$\sum_{i=1}^{k-1} V_i < \sum_{j=k}^{n} V_j \text{, т.е. } V_1+V_2+\ldots+V_{k-1} < V_k + V_{k+1} + \ldots + V_n$$

$$\sum_{i=1}^{k} V_i > \sum_{j=k+1}^{n} V_j \text{, т.е. } V_1+V_2+\ldots+V_k > V_{k+1} + V_{k+2} + \ldots + V_n$$

$$R_z = V_k$$

Якщо існує рівність:

$$\sum_{i=1}^{k} V_i = \sum_{j=k+1}^{n} V_j \text{, т.е. } V_1+V_2+\ldots+V_k = V_{k+1} + V_{k+2} + \ldots + V_n,$$

то існує нескінченна безліч роздільних значень, що містяться в інтервалі між V_k і V_{k+1} в цьому випадку за роздільне значення приймають середню арифметичну

$$R_z=\frac{V_k+V_{k+1}}{2}$$

Приклад. Розглянемо наступний ряд ознак:

2, 4, 5, 6, 8, 10, 12

Маємо такі нерівності:

2 + 4 + 5 + 6<8+10+12; 17 < 30;

2 + 4 + 5 + 6 + 8 > 10 + 12; 25 > 22,

звідки випливає, що $R_z = 8$

Приклад. Розглянемо ряд значень деякої ознаки:

3, 4, 5, 7, 9, 10.

Маємо: 3 + 4 + 5 + 7 = 9 + 10.

Таким чином, за розділове значення можна прийняти будь-яку величину, що міститься у інтервалі між 7 і 9. Однак будемо приймати за розділове значення середню арифметичну чисел, що обмежують цей інтервал 8

$$R_z = \frac{7+9}{2} = 8$$

1.6 Середні неаналітичні зважені

Середні неаналітичними називаються зваженими, якщо вони підраховуються по групах частот. Однією з основних позиційних зважених є мода.

1.6.1 Мода (Переважне значення)

Модою називається таке значення, яке повторюється в досліджуваній групі найбільше раз (є домінуючим значенням). Це визначення не виключає того практично можливого випадку, коли кілька значень мають одну і ту ж максимальну частоту. Однак у більшості випадків одне яке-небудь переважне значення притягує до себе особливо велике число членів ряду.

Приклад :

Клас	100-119	120-139	140-159	160-179
Частота	2	20	60	15

У цьому розподілі найбільш численним є третій клас (140-159) з частотою 60. Цей клас називають модальним.

Точне значення моди можна отримати за такою формулою:

$$M_0 = W_н + k \cdot \left(\frac{f_M - f_{M-1}}{2 \cdot f_M - f_{M-1} - f_{M+1}} \right),$$

де M_0 - мода, $W_н$ - початок модального класу, k - величина класового проміжку, $f_{м-1}$ - частота класу, що передує модальному, $f_м$ - частота модального класу, $f_{м+1}$ - частота класу, наступного за модальним.

Для наведеного розподілу $W_н = 140$, $k = 10$, $f_{м-1} = 20$, $f_м = 60$, $f_{м+1} = 15$. Отже, мода цього розподілу дорівнює:

$$M = 140 + 10 \cdot \left(\frac{60 - 20}{2 \cdot 60 - 20 - 15} \right) = 144{,}7$$

Зазвичай, якщо класи взяті не занадто дрібні (10-12 класів на всю групу), є всього один модальний клас. У деяких розподілах зустрічається два або три модальних класи. Іноді це може бути наслідком того, що в досліджувану групу потрапив різнорідний

матеріал, що відноситься до різних категорій (більш великої і менш великої) по досліджуваному ознакою.

2. ПОКАЗНИКИ РІЗНОМАНІТНОСТІ ОЗНАКИ

Всяка група складається з особин, що відрізняються один від одного по кожному з ознак. Відмінності ці іноді дуже великі, іноді вони майже непомітні, але вони завжди є, оскільки неможливо знайти навіть двох абсолютно однакових особин.

При вивченні загальних властивостей сукупностей неможливо обмежитися одними середніми величинами, потрібно додатково залучити і такі показники, які характеризували б ступінь різноманітності особин в групі. Такими показниками є:

- ліміти (lim) - максимальне (max) і мінімальне (min) значення;
- середньоквадратичне відхилення (σ);
- коефіцієнт варіації (CV).

Крім того, іноді вживається квартильне і децільне відхилення.

Загальною властивістю показників різноманіття є їх здатність характеризувати різну ступінь і різні особливості різноманітності.

2.1 Ліміти

Найпростішим показником різноманітності групи є ліміти ознаки, тобто наявні максимум і мінімум. Іноді разом з лімітами вказується і розмах ознаки - різниця між максимальним і мінімальним значеннями.

Зазвичай, розмах приписується до лімітів в дужках: 2-7 (5).

Приклад. При вивченні ваги биків у двох господарствах (а) і (б) отримані наступні дані:

а) 640, 645, 650, 655, 660 M = 650

б) 600, 630, 670, 680, 700 M = 650

Середні живі ваги биків в обох господарствах однакові - 650 кг, однак, як видно різноманітність биків за вагою в другому господарстві більше, ніж у першому.

У цьому випадку, найбільш просто показати розмаїття можна за допомогою лімітів і розмаху:

а) lim_1 = 640 — 660 (20);

б) lim_2 = 600 —700 (100).

Як видно, виявилося, що в другому господарстві розмах ваги биків в п'ять разів більше, ніж у першому.

При проведенні паралельних аналізів ліміти отриманих результатів та їх розмах служать показниками якості проведеної роботи. Крім показань ступенем різноманітності, ліміти дають характеристику, як досягнень, так і недоліків, наявних в групі по досліджуваному ознакою.

Приклад. Припустимо, що порівнюються дві групи-яких особин по довжині.

1 2 3 4 5 6 7 8 9

а) 10, 11, 12, 13, 14, 15, 16, 17, 18 M=14, lim=10-18 (8)

б) 10, 14, 14, 14, 14, 14, 14, 14, 18 M=14, lim=10-18 (8)

Середні та ліміти в обох групах однакові, і в той же час ступінь різноманітності цих груп явно різна. У першій групі всі особини різні, а в другій сім особин з дев'яти мають

один і той же розмір. Мінливість першої групи явно більше, ніж другої, але відзначити це за допомогою лімітів в даному випадку неможливо.

У таких і подібних їм випадках, найбільш точно охарактеризувати ступінь різноманітності можна за допомогою особливого показника - середнього квадратичного відхилення.

2.2 Середнє квадратичне відхилення

Середнє квадратичне відхилення має зовсім виняткове значення в математичній статистиці і біометрії, зокрема. Цей показник використовується як абсолютного заходу різноманітності і, крім того, покладений в основу майже всіх характеристик мінливості, розподілу, кореляції, регресії, дисперсійного аналізу.

Середнє квадратичне відхилення визначається за формулою:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(V_i - M)^2}{n-1}},$$

де ν = n - 1 - число ступенів свободи; M - середня арифметична (чи інша середня).

Під коренем - сума квадратів центральних відхилень. За допомогою середньоквадратичного відхилення визначається ступінь різноманітності особин в групі по досліджуваному ознакою.

2.3 Число ступенів свободи

Число ступенів свободи дорівнює числу елементів вільної різноманітності.

Воно дорівнює числу всіх наявних елементів вивчення без числа обмежень різноманітності.

Приклад. Для дослідження потрібно взяти три об'єкти з будь-яким розвитком досліджуваного ознаки. У даному випадку величина ознаки не має жодних обмежень, тому число ступенів свободи одно: ν = 3 – 0 = 3.

Якщо для дослідження береться три об'єкти, але з умовою, що сума значень досліджуваної ознаки повинна дорівнювати певній величині, наприклад, 100, то перший об'єкт може мати ознаку будь-якої величини, наприклад, 20 (перша ступінь свободи), другий об'єкт може також мати будь-яке значення ознаки, наприклад, 30 (друга ступінь свободи), третій же об'єкт може мати тільки одне певне значення 50 (50 = 100 - 20 - 30) і тому не має свободи різноманітності.

Таким чином, для трьох дат при одному обмеженні (умові) різноманітності є два ступеня свободи (ν = 3 - 1 = 2).

Для n дат при k обмеженнях мається ν = n - k ступенів свободи. Наприклад, при обчисленні середньої арифметичної вся сума значень ознаки відноситься до одного елементу з числа, що утворюють цю суму, причому ніяких обмежень величини значень ознаки немає. Тому число елементів вільної різноманітності, що утворюють середню арифметичну, дорівнює числу дат.

При обчисленні середнього квадратичного відхилення є одне обмеження величини ознаки у досліджуваних об'єктів. Сигма обчислюється для певної групи, що має певну середню арифметичну. Тому різноманітність елементів, що утворюють середньоквадратичне відхилення, обмежена цією однією умовою і в даному випадку число ступенів свободи дорівнює числу дат без однієї.

2.4 Коефіцієнт варіації

Середнє квадратичне відхилення є основним показником різноманітності дат, що об'єднуються в досліджувані групи. При цьому сигма служить безпосереднім показником різноманітності тільки при дотриманні наступних умов:

1) порівнюються тільки однакові ознаки;

2) середні порівнюваних груп не повинні сильно (< 5 %) відрізнятися один від одного.

Наприклад, якщо для довжини дзеркального коропа в одному улові $M_1 = 28$ см і $\sigma_1 = 2$ см, а в другому улові $M_2 = 27$ см , і $\sigma_2 = 5$ см, то ясно, що в другому улові різноманітність більше і риби менш стандартні.

Якщо зазначені умови не виконуються і необхідно порівнювати різноманітність різних ознак або однакових ознак при різкому розходженні середніх, сигма безпосередньо не може бути використана для порівняння різноманітності.

Приклад . Є дані про величину середнього квадратичного відхилення таких ознак:

- жива вага при народженні 3 кг;
- відсоток жиру в молоці 0,2 %;
- жива вага дорослих корів 48 кг;
- висота в холці 7,2 см;
- удой за лактацію 600 кг.

За цими даними неможливо встановити, який із зазначених ознак більш різноманітний. Не можна порівняти 600 кг удою з 7,2 см висоти в холці або з 0,2 % жиру і т.д.

У цьому випадку для порівняння різноманітності різних ознак застосовується особливий показник - коефіцієнт варіації CV. Цей показник є функцією обох основних показників - середнього квадратичного відхилення і середньої арифметичної, виражається абстрактним (безрозмірним) числом і тому дуже зручний для порівняння різноманітності будь-яких признаків. Обчислюється коефіцієнт варіації за такою формулою:

$$CV = \frac{\sigma}{M} \cdot 100\%$$

Наприклад, якщо $\sigma = 30$ і $M = 150$, то:

$$CV = \frac{30}{150} \cdot 100\% = 20\%$$

Визначення коефіцієнта варіації для наведеного вище прикладу вносить достатню ясність у питання про те, яка з ознак більш різноманітна (див. табл.2.1).

Таблиця 2.1

Признак	M	σ	CV
Жива вага при народженні	30 кг	3 кг	10 %
Жива вага дорослих корів	400 кг	48 кг	12 %
Удой за лактацію	3000 кг	600 кг	20 %
Відсоток жиру в молоці	4,0 %	0,2 %	5 %
Висота в холці	120 см	7,2 см	6 %

Виявилося, що у дослідженої групи тварин найбільш різноманітною, мінливою ознакою є удій за лактацію, а найменш мінливою - жирномолочність. Висота в холці

менше мінлива, ніж жива вага, а жива вага при народженні трохи менш мінлива, ніж жива вага дорослих корів.

2.5 Нормоване відхилення

Зазвичай ступінь розвитку ознаки визначається шляхом його вимірювання і виражається певним іменованих числом: 3 кг ваги, 15 см довжини, 4 % жиру в молоці, 15 кг настригу вовни, 700 г приросту ваги на добу та ін. Цей основний спосіб характеристики ознак виявляється недостатнім, коли потрібно ще і оцінити отримане значення, тобто визначити, чи можна його вважати значним або, навпаки, недостатнім, або перебувають в нормі, вибрати краще і т.д.

Припустимо, що з двох корів треба вибрати одну, кращу по удою. Перша дала за 300 днів лактації 3500 кг молока, друга в тому ж господарстві за той же рік дала 4500 кг за 300 днів лактації.

Чи можна на підставі лише цих даних зробити висновок, що другий корова краще перше по обільномолочності? Ні, ще не можна. При всіх інших рівних умов (оптимальні умови годування й утримання, приблизно рівні періоди сухостою, дати отелення, тривалості лактації і т. д.) корови змінюють свій удой залежно від віку, тощо.

Для отримання повних оцінок виміряних значень ознак прийнятий спеціальний показник - нормоване відхилення, який розраховується за формулою:

$$x = \frac{V - M}{\sigma},$$

де x - нормоване відхилення; V - дата, результат безпосереднього вимірювання ознаки; M - середня арифметична відповідної групи, з якої взята досліджувана особина; σ - середньоквадратичне відхилення цієї ознаки в групі.

Таким чином, нормоване відхилення показує, на скільки сигм відхиляється значення ознаки від середньої для відповідної групи.

Нормоване відхилення - величина неіменована, що представляє велику зручність при порівнянні розвитку різних ознак. За допомогою нормованого відхилення можна вести порівняльну оцінку особин, які належать до різних видів, різних порід, віків, за різними ознаками.

3 ЗАКОНИ РОЗПОДІЛУ ОЗНАКИ У ВИБІРКАХ

Різноманітність об'єктів складових групу (вибірку) це основна властивість всякої сукупності. У нечисленних групах важко визначити яку-небудь закономірність в різноманітності дат. У міру збільшення чисельності досліджуваних груп все більше проявляються закономірності в їх різноманітності, які приховані (непомітні) в нечисленних групах.

Якщо є численна група особин, то різні значення ознаки зустрічаються в цій групі різне число разів. Це явище називається розподілом ознаки.

При вивченні еколого-біологічних об'єктів за різними ознаками можна зустріти кілька типів розподілу ознаки в досліджуваній групі. У біометричних дослідженнях найбільше значення мають такі закони розподілу:

нормальне;

биномиальное;

рідкісних подій (Пуассона).

Зобразити розподіл ознаки можна наступними основними способами:

варіаційним рядом;

варіаційної кривої;

гистограмою;
кумулятою.

Варіаційний ряд - це впорядковане відображення реально існуючого розподілу значень ознаки, по окремих особинам вивченої групи. Варіаційний ряд представляє собою подвійний ряд чисел, що складається з позначень класів і відповідних їм частот.

Приклад. Необхідно побудувати варіаційний ряд для 1000 дат по 11-ти класах, через 20 одиниць починаючи з 110.

Мода (зустрічається найбільш часто f_{max}= 250)

Середини класів (W)	110	130	150	170	190	210	230	250	270	290	310	Сума
Частота (f)	2	20	60	160	250	240	180	70	15	2	1	1000

Медіана

Розмах = Max – Min = 310 – 110 = 200.

Варіаційний ряд включає в себе весь первинний матеріал з вимірювання однієї якого-небудь ознаки у всіх представників досліджуваної групи. Це дозволяє привести експериментальний матеріал в певний порядок і для дуже численних груп визначити показники, що характеризують ознаку, як за середнім рівнем розвитку, так і по деталях різноманітності.

Варіаційний ряд дозволяє без конкретних обчислень визначити величину середнього рівня ознаки і різноманітності.

3.1 Складання варіаційного ряду

При складанні варіаційного ряду всі величини ознаки розбиваються на рівні інтервали - класи. Попередньо необхідно встановити:

число класів;
величину класів;
границі класів;
середини класів;
частоти за класами.

Число класів. Весь розмах значень ознаки від мінімуму до максимуму розділяється зазвичай на 8-12 рівних інтервалів. При точних дослідженнях число класів встановлюється за такою формулою:

$$R = 1+ 3{,}3\cdot\lg(n),$$

де n - число дат; lg (n) - логарифм десятковий від числа дат, наприклад: lg(100)=2.

Після обчислення за цією формулою величину R округлюють у більшу сторону до найближчого цілого числа.

Величина класів або величина класового проміжку дорівнює розмаху значень від мінімуму до максимуму, поділеній на неокруглене число класів R.

Зазвичай величина класів встановлюється за формулою:

$$k= \frac{V_{max}-V_{min}}{1+3.3\cdot\lg(n)}$$

де Vmax - максимальне значення, Vmin - мінімальне значення.

Отримане дробове число при діленні округлюють до найближчого цілого числа. Наприклад, якщо отримано 43,4 то за величину класового проміжку k потрібно взяти 44.

Межі класів. Кінець кожного класу повинен бути менше початку наступного на величину, рівну прийнятої точності вимірювання (ξ).

$$W_{в(1)} = V_{max} + 0{,}5\ k - \xi \qquad W_{н(1)} = V_{max} - 0{,}5 \cdot k,$$

для i = 2 . . R:

$$W_{в(i)} = V_{в(i-1)} - k \qquad W_{н(i)} = V_{н(i-1)} - k$$

Наприклад, якщо вимірюється довжина тварин з точністю до ξ = 1 см і встановлена величина класового проміжку 5 см, то межі класів, починаючи з нижнього мінімального будуть такими: 100-104, 105-109, 110 -114, 115-119 і т.д.

Середини класів встановлюються двома способами. Якщо ознака може бути виражений будь-яким числом - і цілим і дробовим, то для встановлення середини класу потрібно до початку класу додати половину класового проміжку.

У тих випадках, коли ознака виражається тільки цілими числами, середина класів дорівнює напівсумі початку і кінця класу.

Частоти класів встановлюються шляхом розноски дат по класах. Позначаються частоти класів символом f. Кожна дата, потрапивши у відповідний клас, прирівнюється за величиною до всіх інших дат, що потрапили в цей клас.

3.2 Гістограма

Гістограма - варіаційний ряд представлений у вигляді діаграми (рис. 3.1), в якій різна величина частот зображується різною висотою стовпців. На гістограмі наочно виявляються особливості розподілу.

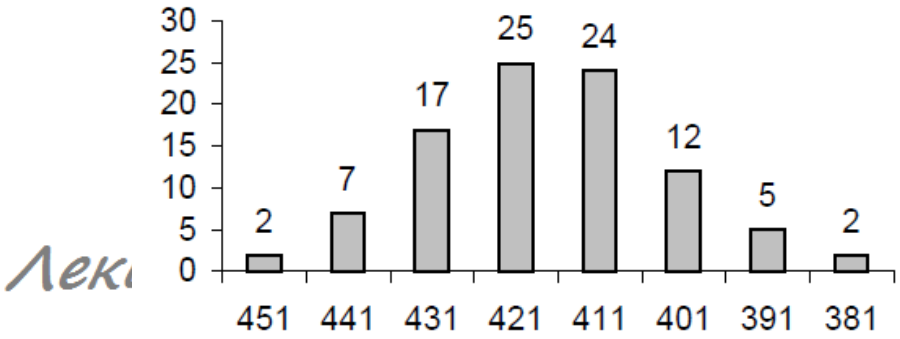


Рисунок 3.1 – Гістограма

3.3 Варіаційна крива

Варіаційна крива - це зображення варіаційного ряду у вигляді кривої (рис. 3.2), ординати якої пропорційні частотам варіаційного ряду. Варіаційна крива - це зручний і наочний спосіб ілюстрації варіаційного ряду в тих випадках, коли на одному графіку потрібно розташувати або зобразити декілька розподілів.

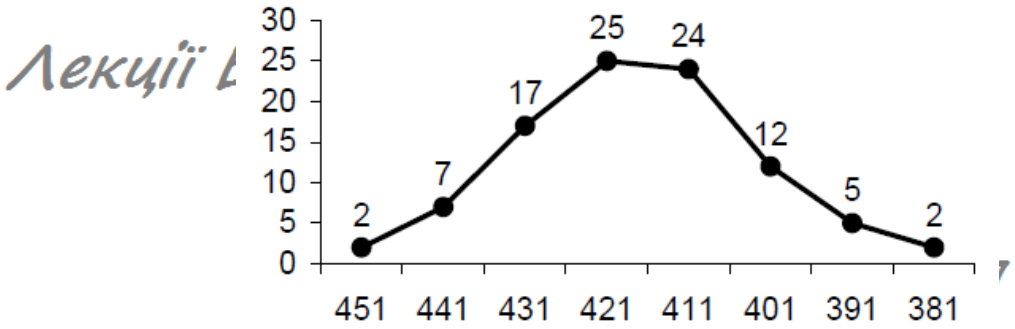


Рисунок 3.2 – Варіаційна крива

Приклад. Потрібно обробити результати дослідів, в яких насіння помідорів піддавалися опроміненню різними дозами рентгенівських променів: 2х, 4х і 8-ми кратномими від норми. На контрольному (висіяні неопромінені насіння) і трьох дослідних ділянках, на випадково обраних 100 кущах рослин підраховувалася число зав'язалися плодів. Результати розподілу кущів (частот) по числу плодів, що зав'язалися (2-4-6 ... 20-22) для неопромінених посівів і посівів з трьома різними дозами опромінення (2, 4 ,8) наведені в табл.3.1 і на рис. 3.3 нижче.

Таблиця 3.1 – Результати розподілу кущів

Клас	1	2	3	4	5	6	7	8	9	10	11
W	2	4	6	8	10	12	14	16	18	20	22
0р	-	-	5	22	45	19	7	2	-	-	-
2р	-	-	4	18	42	25	8	3	-	-	-
4р	-	1	1	2	2	12	21	40	11	8	2
8р	5	33	52	8	2	-	-	-	-	-	-

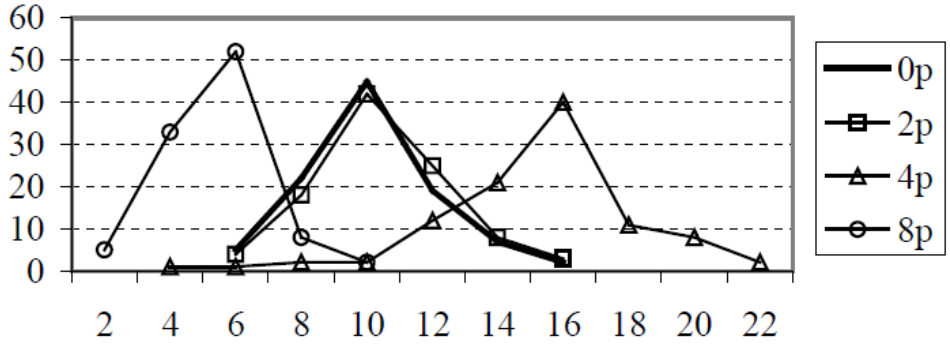


Рисунок 3.3 - Результати розподілу кущів

Зіставлення чотирьох варіаційних кривих дозволяє зробити наступний висновок:

доза 2р істотно не збільшує проти контролю ні середнього числа плодів, ні різноманітності цієї ознаки;

доза 4р надає явно підвищену дію і на середній рівень, і на різноманітність (збільшується число рослин з підвищеним рівнем плодів, що зав'язалися: 14-22);

доза 8р пригнічує утворення плодів.

3.4 Кумулята

Кумулята - це зображення розподілу ознаки у вигляді кривої (рис. 3.4), ординати якої пропорційні накопиченим частотах варіаційного ряду. Щоб скласти ряд накопичених

частот, потрібно до частот найменших класів додати частоту наступного класу, тобто визначити накопичення суми частот по всіх класах.

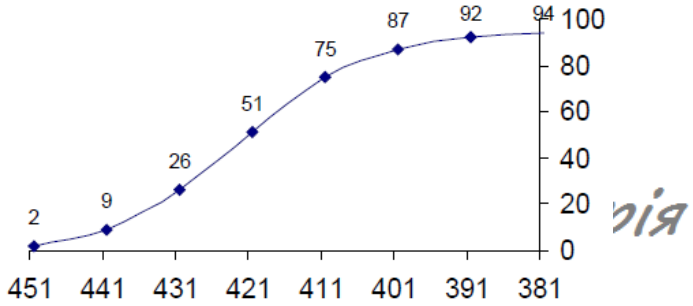


Рисунок 3.4 - Кумулята

Кумулята має перевагу перед варіаційною кривою в разі вивчення процесу накопичення якої-небудь ознаки. Один з цих методів - метод накопичення - показує деталі дії отрут і отруйних речовин у живих організмах і в природі в цілому.

3.5 Нормальний розподіл

У більшості розподілів, з якими доводиться зустрічатися при вивченні природних явищ і об'єктів екологові, біологу, рослинникам, зоотехніку, медичному працівнику, проявляється певна закономірність:

1. Крайні значення - найменше та найбільше - з'являються рідко;
2. Чим ближче значення ознаки до середньої арифметичної, тим воно частіше зустрічається;
3. У центрі розподілу є такі значення, які зустрічаються найбільш часто і утворюють в варіаційному ряду модальний клас.

Даний розподіл значень ознаки так часто зустрічається в самих різних галузях науки і практики, що спочатку воно приймалося за норму всякого масового випадкового прояву ознак і відповідно з цим отримало особливу назву - нормальне.

В даний час нормальним називають розподіл, який з достатнім для практики наближенням слідує закону, відкритому трьома вченими в різний час: Муавром в 1733 р. (Англія), Гауссом в 1809 р. (Німеччина) і Лапласом в 1812 р. (Франція).

Закон нормального розподілу виражається наступною формулою:

$$p^* = \frac{n \cdot k}{\sigma} \cdot \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{x^2}{2}\right) \qquad (3.1)$$

де: p * - теоретична частота кожного класу розподілу;
n - обсяг групи, число об'єктів дослідження;
k - класовий проміжок (величина класів);

σ - середнє квадратичне відхилення: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(V_i - M)^2}{n-1}}$

$x = \frac{W - M}{\sigma}$ — нормоване відхилення середин кожного класу розподілу.

Співвідношення у формулі (3.1): $\frac{1}{\sqrt{2\cdot\pi}}\cdot\exp\left(-\frac{x^2}{2}\right)$ є f (x) - функція нормованого відхилення, яку можна розрахувати для будь-яких значень x.

У загальному випадку, для визначення виду розподілу (нормальне, біноміальне або Пуассона) досліджуваної ознаки виконують зіставлення емпіричних і теоретичних частот цього розподілу між собою.

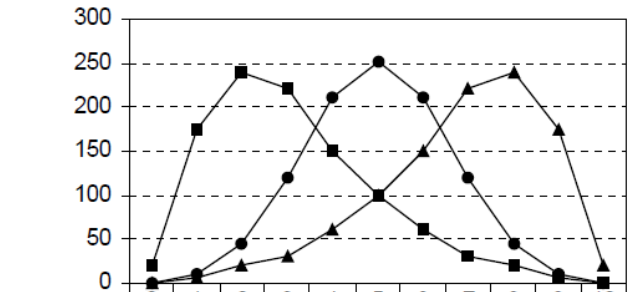
Знаходження ряду теоретичних частот для наявного емпіричного розподілу називається вирівнюванням емпіричних кривих по нормальному або іншому закону, яке буде розглянуто далі. Цей процес має дуже велике теоретичне і практичне значення. Вирівнювання емпіричних кривих розкриває закономірність розподілу, яка зазвичай прихована під випадкової формою свого прояву.

Наступною важливою властивістю будь-якого розподілу, в тому числі і нормального, є те, що можна передбачити ймовірність появи такого значення ознаки, яке знаходиться в межах заданих меж, віддалених в обидві сторони від середньої на будь-яке число сигм (середніх квадратичних відхилень).

3.5.1 Асиметрія і ексцес

Деякі ознаки у рослин і тварин при об'єднанні цих об'єктів в групи дають розподіл, що значно відрізняється від нормального.

У тих випадках, коли які-небудь причини сприяють появі значень ознаки, що відрізняються від середньої величини в бік зменшення або в бік збільшення, утворюються асиметричні розподілення. Відповідно до цього розрізняють ліву і праву асиметрії (див. рис. 3.5).



	0	1	2	3	4	5	6	7	8	9	10
Ряд1	1	10	45	120	210	252	210	120	45	10	1
Ряд2	20	175	240	220	150	100	60	30	20	7	1
Ряд3	1	7	20	30	60	100	150	220	240	175	20

Рисунок 3.5 - Асиметрія

У тих випадках, коли які-небудь причини сприяють переважному появі і середніх і крайніх значень ознаки, утворюються позитивні ексцесивні розподіли, що мають вид гострої піраміди з розширеною підставою (рис. 3.6).

При негативному ексцесі в центрі розподілу є не вершина, а западина, розподіл стає двумодальним, варіаційна крива - двувершиною .

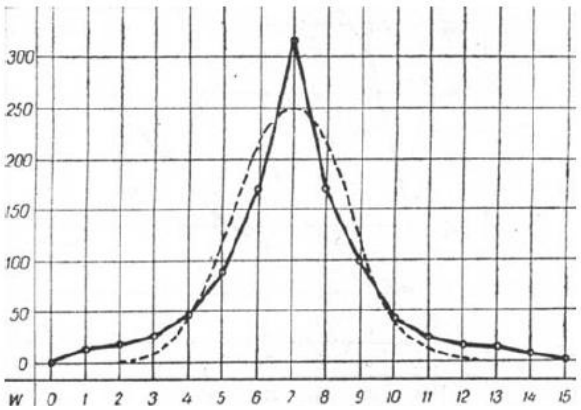


Рисунок 3.6 – Ексцес

3.6 Достовірність відмінності розподілів

Часто для практичних і наукових робіт необхідно встановити наскільки сильно або слабо розходяться між собою емпіричні та теоретичні ряди. Іншими словами, виникає необхідність встановити таку межу, досягнення якого означає, що розбіжність між емпіричним і теоретичним (нормальним, біноміальним і т.д.) розподілом є настільки великим, що з ним необхідно рахуватися і даний емпіричний ряд не можна приймати, в даному випадку, за нормальний (біноміальний або інший).

У екологічних (біометричних) дослідженнях для цієї мети застосовуються особливі показники - критерій χ2 (хі-квадрат) і критерій λ (лямбда).

3.6.1 Критерій χ2 (хі-квадрат, Пірсона)

Критерій χ2 запропонований Пірсоном і застосовується у всіх випадках, коли необхідно визначити ступінь відхилення фактичного розподілу частот від теоретичного.

Визначається величина χ2 за такою формулою:

$$\chi^2 = \sum_{j=1}^{m} \frac{(f_j - p^*)^2}{p_j^*},$$

де : m - число класів; f - емпірична частота; p * - теоретична частота.

Якщо крайні класи розподілу мають теоретичні частоти менше одиниці, то при обчисленні χ2 їх попередньо необхідно об'єднати в один клас разом з найближчим класом, що має частоти p* > 1. Разом з теоретичними треба об'єднати і відповідні фактичні частоти.

Після знаходження величини χ2 потрібно визначити, велика або мала вона для даного розподілу. Для цього користуються таблицею граничних (стандартні) значень $\chi 2_{st}$.

Для того, щоб користуватися цією таблицею, необхідно попередньо встановити число ступенів свободи для досліджуваного розподілу. Якщо як теоретичного розподілу береться нормальне, всі деталі якого визначаються двома постійними величинами M і σ, то число ступенів свободи в таких випадках дорівнює числу класів без двох. За число класів береться те, яке вийшло після об'єднання класів з дробовими теоретичними частотами.

$$\chi^2 = \sum_{j=1}^{m} \frac{(f_j - p^*)^2}{p_j^*} \geq \chi^2_{st}$$ { b_1 - для малої ; b_2 – для обичної; b_3 – для великої

відповідальності дослідження

При виконанні цієї умови розбіжність вважається достовірним і спостережуваний розподіл не можна вважати відповідним прийнятому спочатку теоретичним розподілом (нормальне, біноміналне або інше).

Для кожного числа ступенів свободи вказані три цифри граничних значення $\chi 2_{st}$, що відповідають трьом стандартним ступенями ймовірності (b_1 = 0,95, b_2 = 0,99 і b_3 = 0,999) того, що розподіли, що показали такі значення $\chi 2$ чи великі, розрізняються достовірно. Під достовірним розходженням розуміється така розбіжність розподілів, яке не може відбутися в порядку звичайних випадкових відхилень фактичних частот від теоретичних.

Якщо фактичне значення $\chi 2$ більше третього значення, відповідного b_3 = 0,999, то у всіх випадках можна вважати відмінність між розподілами достовірним.

3.6.2 Критерій λ (лямбда)

Критерій λ запропонований вченими О.М. Колмогоровим і Н.В. Смирновим і може застосовуватися для визначення достовірності розбіжності між фактичними і теоретичними розподілами, а також відмінностей між будь-якими двома розподілами частот однієї і тої ж ознаки навіть у тому випадку, коли число класів і число дат у цих розподілів неоднаково. Для застосування критерію λ не вимагається визначати число ступенів свободи і не потрібні таблиці для визначення трьох граничних значень критерію, так як для будь-якого числа класів ці граничні значення однакові: 1,36; 1,63; 1,95 і відповідають звичайним трьома ступенями ймовірності достовірного відмінності - b_1 = 0,95 ; b_2 = 0,99; b_3 = 0,999.

Єдиною умовою застосування критерію λ є достатня чисельність порівнюваних розподілів - не менше кількох десятків, а краще сотень дат. Для порівняння емпіричного розподілу з теоретичним при однаковому числі класів і при однаковій загальній, чисельності груп критерій ламбда визначається за наступною формулою:

$$\lambda = \frac{|d|}{\sqrt{n}} = \frac{\left|\sum_{i=1}^{m} f - \sum_{i=1}^{m} p_i^*\right|_{\max}}{\sqrt{n}}$$

де d - максимальна абсолютна різниця (без урахування її знаку) між накопиченими частотами в емпіричному і теоретичному розподілах для одного і того ж класу; n - загальне число дат, що утворили емпіричний розподіл.

Для визначення критерію лямбда потрібно скласти ряди накопичених частот (кумуляти) для обох порівнюваних розподілів $\sum f_i$ та $\sum p_i^*$, далі взяти найбільшу різницю (без урахування її знаку) між цим величинами і отриману різницю розділити на $\sqrt{n}$.

Приклад. При дослідних посівах нового сорту пшениці поле було розбито на 840 ділянок. З поділяночних врожаїв було складено розподіл з класами через 7 г/м^2. Потрібно перевірити, чи можна вважати отриманий розподіл нормальним.

Для цієї мети необхідно скласти теоретичний нормальний розподіл за отриманими значеннями M та σ (див. табл.3.2).

В ряду різниць (d) між накопиченими частотами за обома розподілам найбільшою величиною є 9,8. Це значення береться як чисельник дробу, в якій знаменник дорівнює кореню квадратному з 840:

$$\lambda = \frac{9.8}{\sqrt{840}} = 0.34 < 1.36$$

Величина λ менша першого граничного значення (1,36) - вказує на те, що розбіжність між фактичним і теоретичним розподілами недостовірно і розподіл врожаю пшениці по ділянках можна вважати нормальним.

Таблиця 3.2 - Данні

№	W, г/м^2	f	p*	$\sum f_i$	$\sum p_i^*$	$\lvert d \rvert$
1	70	1	0.1	1	0.1	0.9
2	77	4	0.9	5	1.0	4.0
3	84	7	6.0	12	7.0	5.0
4	91	19	25.2	31	32.2	1.2
5	98	72	73.6	103	105.8	2.8
7	105	141	147.1	244	252.9	8.9
4	112	201	201.9	445	454.8	**9.8**
8	119	203	189.4	648	644.2	3.8
9	126	125	121.2	773	765.4	7.6
10	133	54	53.7	827	819.1	8.9
11	140	9	16.4	836	835.5	0.5
12	147	2	3.4	838	838.9	0.9
13	154	1	0.5	839	839.4	0.4
14	161	1	0.1	840	839.5	0.5

3.7 Біноміальний розподіл

Група особин може вивчатися не тільки за кількісними ознаками, які можуть мати різну ступінь свого прояву і вимірюватися іменованими величинами - в кілограмах, літрах, сантиметрах і інших одиницях виміру. Є ознаки, які зазвичай не мають кількісних градацій (чоловіча стать, червона масть та ін.). У кожної окремої особини така ознака може бути присутня або відсутня. Такі ознаки називаються якісними або альтернативними.

Принципової різниці між кількісними і якісними ознаками немає. У більшості ознак, які вважаються якісними, при більш ретельному вивченні може бути знайдена і виміряна ступінь його прояву, і тоді якісна ознака розглядають як кількісна.

Характеристика групи за якісним ознакою полягає у вказівці того, скільки в цій групі є особин з наявністю даної ознаки і у скількох особин його немає. Для такої характеристики вживаються такі позначення необхідних параметрів:

n - загальна кількість особин в групі (наприклад, 200);

i - кількість особин, що мають досліджувану ознаку в групі (120);

j - кількість особин, що не мають даної ознаки в групі (80), (j = n - i);

p = $\frac{i}{n}$ – частка особин, що мають ознаку (120/200 = 0,60);

q = $\frac{j}{n}$ – частка особин, що не мають ознаку (80/200 = 0,40);

$r_\sum = \sum_{i=0}^{n} r_i$ загальна кількість досліджених груп особин.

r_i - число груп, що мають і-ту кількість особин, які володіють аналізованою ознакою. Очевидно наступна рівність:

$$p+q=1=\frac{\sum_{i=0}^{n} r_i}{r_\Sigma}$$

Приклад. У кожному десятку з $r_\Sigma = 20$ десятків виловлених риб можуть зустрітися i = {0 (жодної), 1, 2, 3, 4, 5, 6, 7, 8, 9 і всі 10} особин, уражених певною хворобою. Таким чином, кілька десятків не матимуть у своєму складі уражених риб, кілька десятків будуть мати тільки по однієї хворої особини, кілька десятків по 2 особини і т.д.

В результаті складеться розподіл (див. табл.3.3) в якому варіаціями будуть величини i-число особин, що мають досліджувану ознаку в окремих рівночисельний приватних групах (число хворих риб в кожному десятку), а частотами r_i - кількість відповідних рівночисельний груп (число десятків).

Таблиця 3.3 - Розподіл

i	0	1	2	3	4	5	6	7	8	9	10
r_i	1	3	4	5	3	2	1	1	0	0	0

Отриманий розподіл називається біноміальним. Така назва пояснюється наступними його властивостями:

у розподілі ознака може мати тільки два варіанти: він є «+» або його немає «-»;

закономірності такого розподілу мають кількісне вираження, пов'язане з коефіцієнтами розкладання бінома Ньютона, який у застосуванні до цього типу розподілів може бути виражений наступним чином:

$$1=\frac{1}{r_\Sigma}\cdot\sum_{i=0}^{n} r_i=(p+q)^n$$

$$1=(p+q)^n=\frac{1}{r_\Sigma}\cdot\sum_{i=0}^{n}\frac{n!}{i!\cdot(n-i)!}\cdot p^i q^{n-i}$$

факторіал чисел дорівнює: 0!= 1; 1!= 1; 2!= 1. 2; 3!= 2!.3; 4!= 3!.4 і т.д.

де i - кількість особин в групі, які мають досліджувану ознаку; n - загальна кількість особин в групі. У розгорнутому вигляді:

$$r_\Sigma\cdot(p+q)^n=\frac{1}{1}p^0\cdot q^n+\frac{n}{1}\cdot pq^{n-1}+\frac{n\cdot(n-1)}{1\cdot 2}\cdot p^2q^{n-2}+\frac{n\cdot(n-1)(n-2)}{1\cdot 2\cdot 3}p^3q^{n-3}\ldots+\frac{1}{1}\cdot p^nq^0$$

Кожен член бінома може бути представлений у вигляді добутку, з яких перший множник цілком залежить від величини n:

$$f(n)=\frac{n!}{i!\cdot(n-i)!}$$

а другий - від співвідношення p, q і n:

$$f(p,q)=p^i\cdot q^{n-1}$$

Підставляючи в формулу бінома величини f (n) і f (p, q), а потім вирішуючи її відносно величини p, можна отримати такі значення:

$p^0\cdot q^n$ – нульовий член бинома (що містить p^0 в нульової ступеня), дає очікувану частку таких рівночисельний груп, в яких з n особин жодна не має досліджуваного ознаки;

np^1q^{n-1} – перший член бинома (з p^1), дає частку груп, в яких тільки одна особина має очікувану ознаку;

$\frac{n(n-1)}{2}p^2q^{n-2}$ – другий член бинома (з p^2), дає частку груп, в яких досліджуваний ознака має по дві особини;

p^n – останній член бинома, дає частку рівночисельних груп, в які всі n особин мають досліджувану ознаку.

Висновок:

Обсяги вибірок кожної групи не повинні відрізнятися між собою, тобто повинні бути рівно чисельними.

Для вивчення необхідно брати таку кількість груп, яке було б не менше числа особин у кожній рівночисельній групі.

Приклад. Серед деякої популяції квітів семипелюсткова форма квітки зустрічається у 10 % рослин. Якщо взяти випадковим чином 100 груп (букетів) по 5 рослин в кожному з різних місць, то скільки можна чекати букетів без семи пелюсткової форми квіток (0) і букетів з 1, 2, 3, 4 і 5 рослинами, що мають такі квітки.

У даному випадку мається $r_\Sigma = 100$ груп, по n = 5 рослин в кожній, причому загальна частка рослин, що мають досліджувану ознаку (яка була визначена заздалегідь) дорівнює p=0,1. Визначення очікуваних частот r_i такого розподілу представимо в табл. 3.4.

Таблиця 3.4 - Визначення очікуваних частот

i	f(n)	f(p,q)	$r_i=r_\Sigma \cdot f(n) \cdot f(p,q)$
0	1	$p^0q^5 = 1\left(\frac{9}{10}\right)^5 = \frac{59049}{100000}$	59
1	5	$p^1q^4 = \left(\frac{1}{10}\right)\left(\frac{9}{10}\right)^4 = \frac{6561}{100000}$	33
2	10	$p^2q^3 = \left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^3 = \frac{729}{100000}$	7
3	10	$p^3q^2 = \left(\frac{1}{10}\right)^3\left(\frac{9}{10}\right)^2 = \frac{81}{100000}$	1
4	5	$p^4q^1 = \left(\frac{1}{10}\right)^4\left(\frac{9}{10}\right)^1 = \frac{9}{100000}$	-
5	1	$p^5q^0 = \left(\frac{1}{10}\right)^5 1 = \frac{1}{100000}$	-

Розрахунки показують, що при загальній частці p = 0,1 - рослин, що мають даниу ознаку і r_Σ = 100 груп по n=5 рослин в кожній, може зустрітися 59 груп без семи пелюсткової форми квіток, 33 групи, в яких з п'яти рослин одна буде з семипелюстковими квітками, 7 груп, які мають з п'яти дві таких рослини, і 1 група, в якій з п'яти особин три матимуть досліджувану ознаку. Поява груп з чотирма і п'ятьма рослинами, що мають семипелюсткову форму квітки, за даних умов малоймовірно.

Легко підрахувати, що при r_Σ = 1000 можна очікувати, що тільки один букет з тисячі матиме чотири рослини з семи пелюстковими квітками, і тільки взявши 10000 (десять тисяч) таких букетів, можна очікувати, що серед них буде один, в якому всі п'ять рослин будуть такими.

Розглянутий раніше розподіл (і наступний - розподіл Пуассона) можна, з деякою часткою умовності, вважати окремими випадками біноміального розподілу. Наприклад, для нормального p → 0,5 і q → 0,5 (рис. 3.7).

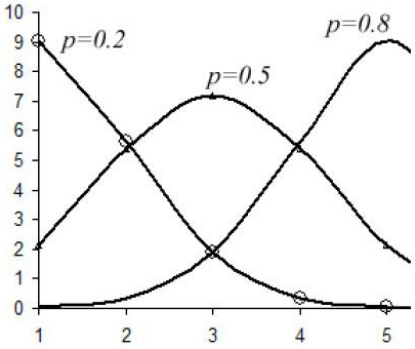


Рисунок 3.7 – Графіки біноміального розподілення

3.8 Розподіл рідкісних подій (Пуассона)

Події, що відбуваються рідко, один або невелике (одиничне) число раз на 1000, 10000 і більше число звичайних явищ, можуть бути зведені в особливий розподіл, в якому варіаціями є різне число рідкісних випадків, а частотами - кількість великих груп, серед яких рідкісна подія відбулося певне число разів.

Цей розподіл можна зіставити з попереднім біноміальним розподілом, у якого $p \to 0$, а $q \to 1$ (см. рис. 3.7).

Розподіл таких рідкісних подій зазвичай підпорядковуються певним законом, який виражається формулою, запропонованою Пуассоном:

$$p_x^* = p \cdot e^{-a} \cdot \frac{a^x}{x!}$$

де p_x^* - теоретична частота розподілу, очікуване число великих груп, серед яких рідкісна подія відбулася x разів;

$p = \Sigma p_x$ - загальна кількість досліджених великих груп;

p_x - фактична (емпірична) частота розподілу. Число великих груп, в яких рідкісна подія відбулася x разів;

x - число рідкісних подій, що відбулися в кожній великій групі; зазвичай x дорівнює невеликому цілому числу: 0, 1, 2, 3 і т.д.;

x! - добуток натуральних чисел від 1 до x (факторіал). Вважається, що факторіал нуля дорівнює одиниці: 0! = 1;

$a = \frac{\sum x \cdot p_x}{p}$ - Середня зустрічальність або середнє число рідкісних випадків на кожну велику групу . Є виваженою середньої арифметичної.

Теоретичне розподіл рідкісних подій має одну особливість: у ньому значення середньої величини приблизно дорівнює квадрату сигми (девіаті). Тому, якщо в інших розподілах основних величин дві - M і σ, то у розподілі рідкісних подій обидві основні величини зведені до однієї - *a* (середньому числу таких подій на кожну велику групу).

З цієї особливості розподілу рідкісних подій випливають два наслідки:

1. все теоретичне розподілення може бути побудоване на підставі тільки однієї середньої - зваженої середньої арифметичної;
2. при визначенні достовірності відмінності теоретичного розподілу від емпіричного за допомогою критерію $\chi 2$ число ступенів свободи дорівнює числу класів без одного.

Приклад. При перевірці засміченості насіння конюшини виявилося, що в кожному навішуванні насіння малася різна кількість насіння березки (небезпечний бур'ян) - від 0 до 3. Для з'ясування причин засміченості потрібно визначити, чи є це звичайним рідкісним явищем, що викликається випадковими обставинами чи ні. Для цієї мети було вироблено

зіставлення 1000 навісок фактичного розподілу ознаки з теоретично очікуваним розподілом рідкісних подій (див. табл.3.5).

Таблиця 3.5 - Теоретично очікуваний розподіл рідкісних подій

Фактичне		Теоретичне	$\chi^2=\frac{(f_i+p_i^*)^2}{p_i^*}$	$d=\left\lvert\sum f_i-\sum p_i^*\right\rvert$
x	f_x			
4	0	1	1	1
3	12	13	0.08	2
2	74	76	0.05	4
1	315	303	0.46	8
0	599	607	0.11	0
Сума =	1000	1000	$\sum=\chi^2=1.70$ N=5-1=4 $\chi^2<\chi^2_{st}\{18.5;13.3;9.5\}$	$\lambda=$ $8/\sqrt{1000}=0.25$ $\lambda<\lambda_{st}$ {1,95; 1,63; 1,36}

де: x - число насіння бур'яну - березки;

fx - фактичне (що спостерігається) число груп;

p_x^* - теоретичне число груп.

Виявилося, що поява в насінні конюшини насіння березки цілком відповідає закономірності рідкісних явищ.

4 РЕПРЕЗЕНТАТИВНІСТЬ (ДОСТОВІРНІСТЬ) ВИБІРКОВИХ ПОКАЗНИКІВ

Під репрезентативністю вибіркових показників розуміють визначення їх достовірності. Зазвичай при будь-якому дослідженні використовується два основні методи:

вивчення всіх особин належать даній групі або виду;

вивчається тільки певним чином обрана частина.

Різниця між цими двома методами полягає в тому, що в 1-му випадку проводиться дослідження всієї генеральної сукупності, в 2-му випадку проводяться вибіркові дослідження.

Генеральна сукупність - це весь масив особин певної категорії. Обсяг генеральної сукупності визначається відповідними вимірами. Якщо вивчається вид диких тварин або рослин, то генеральною сукупністю будуть всі особини цього виду. У цьому випадку обсяг генеральної сукупності буде дуже великим і при розрахунках він приймається за нескінченність (∞).

Іноді обсяг генеральної сукупності доступний для суцільного дослідження. Якщо вивчається невелика сукупність, необхідно визначити середні величини. У цьому випадку генеральна сукупність може бути представлена невеликою кількістю особин, але для всіх досліджень. Якщо генеральна сукупність представляє порівняно невелику кількість особин, то вона характеризується генеральними параметрами.

Вибірка - це група об'єктів, яка відрізняється 3-ма особливостями:

являє собою частину генеральної сукупності;

вона вибирається певним чином, але у випадковому порядку;

вибірка досліджується для характеристики всієї генеральної сукупності.

4.1 Способи відбору об'єктів у вибірку

Існує кілька різних способів відбору об'єктів у вибірку.

Випадковий повторний відбір.

У цьому випадку об'єкти вивчення вибираються з генеральної сукупності, але без попереднього урахування розвитку у них досліджуваних ознак, тобто у випадковому для цієї ознаки порядку. Після відбору кожен окремий об'єкт вивчається і повертається в свою генеральну сукупність. Таким чином, кожен об'єкт може повторно потрапити в іншу вибірку.

Розглянутий спосіб відбору рівносильний відбору з нескінченно великою генеральної сукупності, для якої розроблено основні показники співвідношень між вибірковими і генеральними величинами.

Випадковий бесповторний відбір.

У цьому випадку об'єкти, відібрані випадково не можуть повторно потрапити в дану вибірку. Цей відбір є найбільш поширеним способом організації вибірки. Він рівносильний відбору з великої, але обмеженої генеральної сукупності, що враховується при визначенні генеральних показників за вибірковими.

Механічний відбір.

Проводиться відбір об'єктів з окремих частин генеральної сукупності.

Ці частини попередньо намічаються механічно по квадратах дослідного поля, по випадкових групах тварин, узятих з різних ареалів проживання популяції і т.д.

Зазвичай намічається стільки частин, скільки передбачається взяти об'єктів для вивчення, тому їх число дорівнює чисельності вибірки. Механічний відбір іноді здійснюється вибором для вивчення особин через певне число, наприклад, при пропущенні тварин через розкол і відборі кожного десятого, сотого і т.д., або відборі одного об'єкта через кожні зустрілися 10, 100 і т.д. примірників при дослідженні всієї популяції.

Типовий пропорційний відбір.

Він передбачає необхідність попереднього вивчення генеральної сукупності за загальнобіологічними або господарським особливостям. На основі такого вивчення вся генеральна сукупність розбивається на частини, наприклад, по типу рослинних угруповань, в яких мешкає вид, по рельєфу місцевості і т.д.

З кожної такої частини для вивчення вибирається у випадковому порядку число примірників, пропорційне населеності окремих частин.

Наприклад, при вивченні певної породи риб беруться улови з різних водойм, і при цьому з кожного улову береться число примірників пропорційне ступеня заселеності чи обсягу водойми.

Серійний (гніздо) відбір.

У цьому випадку генеральна сукупність розбивається на частини (серії), деякі з яких досліджуються цілком. Цей спосіб застосовується тоді, коли досліджувані об'єкти рівномірно розподілені або в рівному обсязі, або на рівній території.

Наприклад, при дослідженні зараженості повітря чи води мікроорганізмами для вивчення беруть окремі проби, які піддаються суцільному дослідженню.

Оскільки частина (вибірка) ніколи не може повністю охарактеризувати все ціле, будь-яка характеристика генеральної сукупності на основі вибіркового дослідження завжди буде не точною і буде мати деяку більшу чи меншу помилку.

Помилки, пов'язані з перенесенням результатів, які отримані при вивченні вибірки, на всю генеральну сукупність, називаються помилками репрезентативності.

4.2 Помилки досліджень

При всякому дослідженні є небезпека допустити цілий ряд помилок найрізноманітнішого характеру. Всі ці помилки можуть бути зведені в наступні групи.

А. Загальні помилки, які властиві як суцільному, так і вибірковому дослідженням. До них відносяться:

1. Методичні помилки. Цей клас помилок пов'язаний з наступними діями:

а) застосування порочної методики проведення досвіду (порушення стандартних правил фіксації препаратів та хімічного аналізу, вибір неправильного напрямку дослідження, невідповідного поставленим завданням, та ін);

б) не вирівняність умов проживання для контрольних і дослідних особин.

2. Помилки точності. Цей клас помилок пов'язаний з наступними діями:

а) використання неперевірених і неправильно градуювальних вимірювальних приладів;

б) розрахунки з недостатньою точністю.

3. Випадкові помилки. З ними пов'язані:

а) описки, прорахунки;

б) сплутування дослідних зразків.

Б. Помилки вибіркового дослідження, які властиві тільки вибірковим дослідженням. До них відносяться:

4. Помилки типовості. З ними пов'язані:

а) відбір у вибірку таких об'єктів, які неправильно, односторонньо відбивають властивості генеральної сукупності, наприклад, дослідження тільки видатних особин або тільки середніх або кращих, чи гірших;

б) відбір у вибірку особин, які розвивалися в умовах, різко відмінних від тих, які характерні для всієї генеральної сукупності;

в) при типовому пропорційному відборі - відбір не з усіх частин популяції і без обліку обсягу типових частин;

г) при серійному (гніздовому) відборі - добір не характерною серії або вивчення тенденційно обраних особин в серії.

Всі зазначені категорії помилок викликаються або неправильної методикою дослідження або невмілим і недбалим виконанням роботи. Уникнути їх або звести до мінімуму можливо за допомогою продуманої і ретельно організованою постановці експерименту в дослідженні.

5. Помилки репрезентативності.

При вибірковому дослідженні існує ще один особливий тип помилок, що випливають із самої сутності вибіркового дослідження і мають причиною обставина, що вся генеральна сукупність повинна характеризуватися на підставі вивчення лише її частини - вибірки.

Помилок репрезентативності неможливо уникнути в вибірковому дослідженні навіть при ідеальній організації дослідницької роботи. Проте, вибіркове обстеження може дати точну характеристику генеральної сукупності внаслідок наявності двох сприятливих обставин:

а) величину помилок репрезентативності можна звести до мінімуму певною організацією вибіркового дослідження;

б) розроблені методи, що дозволяють за вибірковими даними визначити можливу величину помилок репрезентативності з тим, щоб враховувати їх при переході від вибіркових показників до генеральних.

Біометрія на основі математичної статистики:

- дає способи визначення помилок репрезентативності (помилок вибіркових показників) - помилки середньої арифметичної *m*, помилки коефіцієнта кореляції m_r та ін.;

- дозволяє розраховувати величину помилок репрезентативності для вибіркових показників. Якщо досліджуються не вибірки, а генеральні сукупності, визначати помилки репрезентативності не потрібно.

- визначати величину помилок репрезентативності слід тільки в тих випадках, коли організація дослідження виключає всі інші види помилок або коли всі вони зведені до мінімуму.

Наприклад, вивчається вага риб, що йдуть косяком, в якому зазвичай попереду - самки, за ними - молодь і ззаду - самці. Якщо у вибірку потрапили риби головним чином з головної частини косяка, то при визначенні середньої ваги для всього косяка буде допущена помилка типовості: у вибірку потрапили особини тільки з однієї частини генеральної сукупності, що відрізняється від інших частин. У даному випадку розрахунок помилок репрезентативності вже не допоможе, так як відбір особин у вибірку проведений неправильно.

4.3 Помилка вибіркової середньої арифметичної

Помилка репрезентативності середньої арифметичної (m) залежить від двох величин: від ступеня різноманітності ознаки (σ) в генеральній сукупності і від чисельності (обсягу) вибірки (n). Припустимо, що різноманітність ознаки у генеральній сукупності дорівнює нулю. Це означає, що всі особини даної сукупності абсолютно однакові. Прикладом може служити колір пера у одноколірних видів птахів. У таких випадках будь вибірка, навіть в один екземпляр дає точну характеристику всієї генеральної сукупності без якої б то не було помилки репрезентативності. Чим більше різноманітність ознаки, тим він більш мінливий, тим більше можливість потрапити на таку вибірку, середня якої сильно відрізняється від генеральної середньої. Таким чином, чим більше розмаїття, тим більше помилка репрезентативності:

$$m \uparrow \approx \sigma \uparrow$$

Легко також зрозуміти залежність помилки вибіркової середньої від чисельності вибірки (n). Чим більше ця чисельність, тим більша частина генеральної сукупності досліджується, тим з меншою помилкою може бути дано висновок про середню для всієї генеральної сукупності.

$$m \uparrow \approx n \downarrow$$

Величина помилки середньої арифметичної визначається в основному виходячи з даних, отриманих у вибірковому дослідженні. Запропоновано декілька формул для розрахунку помилки середньої - для кожного способу відбору об'єктів вивчення.

У більшості біологічних досліджень, при будь-якому способі відбору особин у вибірку можна застосовувати єдину формулу помилки середньої, формулу для випадкового безповторного відбору:

$$m = \frac{\bar{\sigma}}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}$$

де - $\bar{\sigma}$ ступінь різноманітності або вибіркове середньоквадратичне відхилення, отримане для вивченої вибірки; n - величина вибірок; N - обсяг генеральної сукупності.

Отримана за цією формулою величина помилки виявляється злегка завищеною для механічного, типового і серійного методів відбору. Це не становить небезпеки, так як при цьому виходить трохи більш суворий підхід до перенесення вибіркових даних на всю генеральну сукупність.

Множник $\sqrt{1 - \frac{n}{N}}$ при n = 0 - перетворює формулу помилки середньої в формулу для випадкового повторного відбору, а при n = N - звертає величину помилки в нуль.

Множник $\sqrt{1 - \frac{n}{N}}$ впливає в тих випадках, коли у вибірку потрапляє значна частина генеральної сукупності, не менше 30-50 %. Зазвичай, коли у вибірці досліджується не більше 5-10 % особин генеральної сукупності, цей множник настільки

близький до одиниці, що практично не змінює значення помилки середньої і для розрахунку помилки середньої можна застосувати більш просту формулу:

$$m = \frac{\bar{\sigma}}{\sqrt{n}}$$

4.4 Розподіл вибіркових середніх

Уявімо, що з генеральної сукупності взято велике число окремих вибірок так, що всі вони вичерпали всю генеральну сукупність: Σn = N. Кожна з цих вибірок буде мати свою середню арифметичну. Всі ці середні величини не однакові і з них можна скласти розподіл, в яке в якості дат увійдуть вибіркові середні (рис. 4.1).

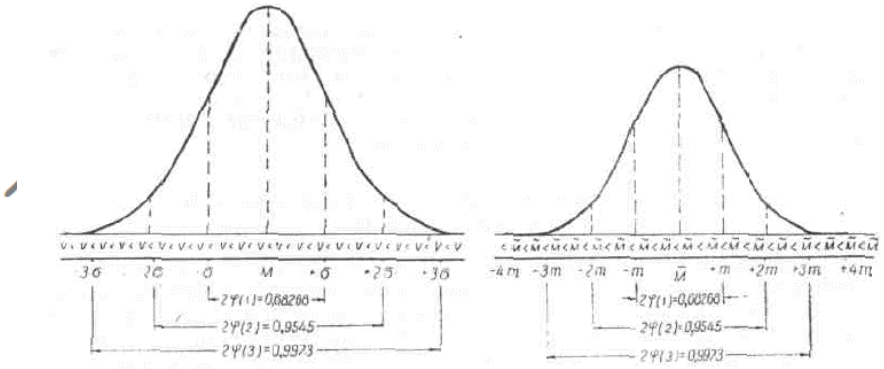


Рисунок 4.1 - Розподіл вибіркових середніх

Як доводиться в математичній статистиці, середня величина цього розподілу буде дорівнює генеральній середній, а середньоквадратичне відхилення цього розподілу вибіркових середніх дорівнюватиме помилці вибіркової середньої, наближену величину якої розраховують за наведеними вище формулами.

Таким чином, розраховуючи величину помилки вибіркової середньої, тим самим визначають з достатньою для практики точністю сигму (σ) ряду, складеного з вибіркових середніх для всіх вибірок із загальної досліджуваної генеральної сукупності.

Дуже важливою властивістю розподілу таких вибіркових середніх є те, що розподіл має достатню близькість до нормального, навіть у тих випадках, коли розподіл індивідуальних дат у генеральній сукупності відрізняється від нормального.

Це означає, що в межах від M $\pm\bar{\sigma}$ міститься 68,3 % дат, які за своїм значенням не відрізняються від середньої величини більше ніж на $\pm\ \bar{\sigma}$. Можливість помилки даного твердження дорівнює 31,7 % або 1 з 3.

Якщо за кордони взяти M ± 2 $\bar{\sigma}$, то виявиться, що відсоток дат дорівнює 95,4 %, які не відміні від середньої не більше ніж на ± 2$\bar{\sigma}$. Тут можливість помилки дорівнює 4,6 % або 1 з 22.

Якщо як кордони значень ознаки взяти M ± 3$\bar{\sigma}$, то виявиться, що всередині цих кордонів у нормальному розподілі міститься 99,7 % дат. Тут можливість помилки дорівнює 0,3 % або 1 з 370. Вона вважається досить малою і в більшості досліджень нею можна знехтувати.

На основі цих міркувань і було введено в практику досліджень «правило трьох сигм», що вважається єдиним критерієм кордонів, умовно включають у себе весь розподіл.

4.5 Три ступеня ймовірності безпомилкового прогнозу при визначенні генеральних величин за вибірковими

Багаторічна і обширна практика застосування методів математичної статистики в екологічних, біологічних та інших дослідженнях показала, що межі допустимих меж за «правилом трьох сигм» занадто великі і виходять з вимог зайвої обережності.

У більшості випадків можна розширити межі, що умовно вміщають весь розподіл, прийнявши за такі межі $M \pm 2\bar{\sigma}$.

В даний час застосовуються три ступені ймовірності того, що укладення про кордони, що вміщають весь розподіл, що не буде помилковим:

$$M \pm 2\bar{\sigma}\ ;\ M \pm 2{,}5\bar{\sigma}\ ;\ M \pm 3\bar{\sigma}$$

чи

$$M \pm 2m\ ;\ M \pm 2{,}5m\ ;\ M \pm 3m$$

Відповідно до цього встановлюються три стандартних фіксованих значення, або порога ймовірності безпомилкового прогнозу:

$$x = \frac{V - M}{y} \qquad t = \frac{\overline{M} - M}{m}$$

поріг t = 2 - імовірність безпомилкового прогнозу допускається для більшості дослідженні в загальній біології, цитології, фізіології, генетиці, ботаніці, зоології, медицині (помилка дорівнює 100 - 95,4 = 4,6 %, тобто 1 з 22).

поріг t = 2,5 - (98,8 %) дає можливість помилитися 1 з 81. Така ймовірність безпомилкового прогнозу потрібно в економічних дослідженнях, пов'язаних з рекомендаціями проведення витрат коштів і праці, а також при обґрунтуваннях реорганізації виробництва.

поріг t = 3 - така ймовірність потрібна в особливо відповідальних роботах: в дослідженнях, перевіряючих спірні теоретичні висновки, в експериментах, що з'ясовують шкідливу дію речовин та ін. (помилка дорівнює 100 - 99,7 = 0,3 %, тобто 1 з 370).

Надалі за показники кордонів, що умовно включають весь розподіл, стали прийматися заокруглені значення ймовірностей для:

1-го ступеня 0,950;

2 -й - 0,990;

3-й - 0,999.

Відповідно до цього були уточнені і значення t для:

1-го ступеня $t_1 = 1{,}96$ (а не 2);

2 -й - $t_2 = 2{,}58$ (а не 2,5);

3-й - $t_3 = 3{,}30$ (а не 3).

Розглянуті ступеня ймовірності справедливі тільки для таких досліджень, які мають справу з досить численними вибірками. При нечисленних вибірках розподіл вибіркових середніх, а також усіх вибіркових величин вже досить сильно відрізняється від нормального і слідує закону розподілу малих вибірок, встановленому англійським ученим Госсетом, який писав під псевдонімом Student (Студент).

Розподіл Стьюдента відрізняється від нормального тим більше, чим менше чисельність вибірки, причому для кожної чисельності малої вибірки мається своє приватне розподілення.

Для кожного значення чисельності малих вибірок можна заздалегідь розрахувати величину порога t для трьох прийнятих ступенів ймовірності.

Наприклад, при 1-му ступені ймовірності (b_1 = 0,95) і при чисельності вибірки n = 10 показник ймовірності t_1 = 2,3, а при чисельності n = 3 - показник t_1 = 4,3.

Значення величини порога t для будь-якої чисельності вибірок і для трьох ступенів ймовірності безпомилкового прогнозу знаходять наближено за формулою:

$$t_\nu = t_\infty \cdot \frac{t_\infty}{\nu + 3 - 1.5 \cdot t_\infty}$$

де t_ν-показник ймовірності для вибірок з числом ступенів свободи - ν;

t_∞ - показник ймовірності для великих вибірок. Залежно від типу відповідальності досліджень він дорівнює або t_1 = 1,96 (b_1 = 0,95), або t_2 = 2,58 (b_2 = 0,99), або t_3 = 3,30 (b_3 = 0,999);

ν - число ступенів свободи. При визначенні генеральної середньої за вибірковою ν=n-1, при визначенні достовірності різниці середніх для некорельованих (не взаємозалежні) вибірок $\nu = n_1 + n_2 - 2$.

5 КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Часто у багатьох дослідженнях потрібно вивчити кілька ознак в їх взаємному зв'язку. Якщо вести таке дослідження стосовно двох ознакам, то можна помітити, що мінливість однієї ознаки знаходиться в деякому відповідно до мінливістю іншої. У деяких випадках така залежність виявляється настільки сильно, що при зміні першої ознаки на певну величину завжди змінюється і друга ознака на певну величину, тому кожному значенню першої ознаки завжди відповідає цілком певне, єдине значення другої ознаки. Такі зв'язки отримали назву функціональні.

Функціональні зв'язку зустрічаються у фізичних і математичних узагальненнях. Наприклад, площа трикутника точно визначається його висотою і підставою, довжина кола - радіусом, швидкість падіння є функцією часу падіння і прискорення сили тяжіння, швидкість протікання певної хімічної реакції знаходиться в залежності від температури.

Необхідно врахувати, що в чистому вигляді функціональні зв'язки зустрічаються тільки в ідеальних умовах, коли передбачається, що ніяких сторонніх впливів немає. На практиці це недосяжно. Ніколи не можна точно виміряти фактично наявний радіус кола, причому обчислена площа ніколи нерівна в точності фактичної, внаслідок практичної неможливості накреслити точну окружність. Швидкість падіння реального тіла в реальних умовах буде завжди різна при одних і тих же часу і прискоренні сили тяжіння. На практиці завжди діють сторонні для даної функціональної залежності фактори, які порушують точність цієї залежності в різних випадках по-різному.

Поки такі порушення залишаються настільки незначними, що їх практично можна не враховувати, зв'язок вважається функціональним.

При вивченні живих об'єктів - диких видів, культурних рослин, домашніх тварин - доводиться мати справу зі зв'язками іншого роду. Живий організм розвивається у зв'язку з умовами його життя, під дією нескінченно великого числа факторів, які по-різному визначають розвиток різних ознак. У живих об'єктів зв'язок між будь-якими двома ознаками настільки часто і сильно порушується і модифікується, що не завжди може бути досить просто виявлений.

У рослин і тварин зв'язок між ознаками зазвичай проявляється особливим чином. Тут кожному певному значенню першого ознаки відповідає не одне значення другої ознаки, а цілий розподіл цих значень при цілком певних основних показниках цього приватного розподілу - середньої величини і ступенем різноманітності. У цьому випадку, такий зв'язок називається кореляційним зв'язком або просто кореляцією.

Кореляційний зв'язок, наприклад, між вагою тварин та їх довжиною виражається в тому, що кожному значенню довжини відповідає певний розподіл ваги (а не одне

значення ваги), таке, що зі збільшенням довжини збільшується і середня вага тварин. Кореляцію класифікують за формою і напрямком, а вимірюють ступенем кореляції.

За формою кореляція може бути:

1) прямолінійною;

2) криволінійною.

У напрямку:

1) прямо-спрямована;

2 обернено-спрямована.

Ступінь кореляції встановлює силу зв'язку між кількісними і якісними ознаками. Вона вимірюється такими показниками:

1) коефіцієнтом кореляції r;

2) кореляційним відношенням;

3) тетрахоричним і поліхоричним показниками зв'язку;

4) приватним і множинним коефіцієнтами кореляції.

Зобразити кореляційний зв'язок двох ознак можна трьома способами (див. рис. 5.1), а саме за допомогою:

1) кореляційного ряду, що складається з ряду пар значень ознак;

2) кореляційної решітки, в якій кожної особині відповідає певна клітина;

3) лінії регресії вісі координат, для якої пропорційні значенням ознак.

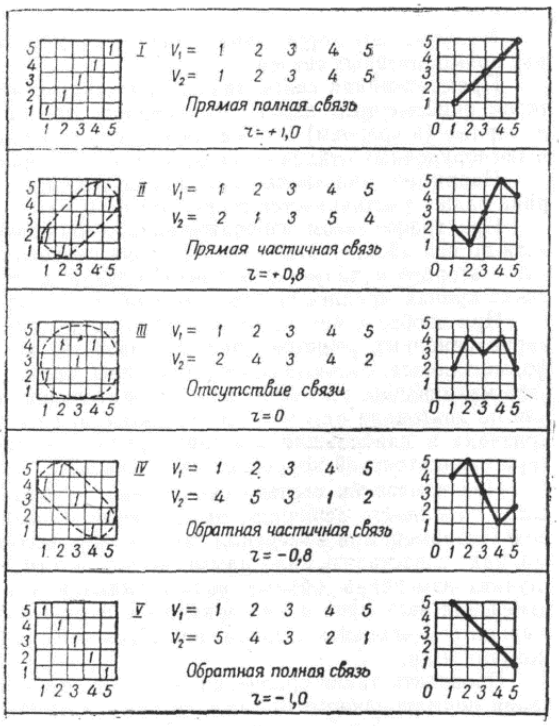


Рисунок 5.1 – Кореляційні зв’язки

5.1 Коефіцієнт кореляції

Коефіцієнт кореляції вимірює ступінь і визначає напрямок прямолінійних зв'язків.

Прямолінійний зв'язок між ознаками, це такий зв'язок, при якому рівномірним змінам першої ознаки відповідають рівномірні (в середньому) зміни другої ознаки.

Наприклад, при збільшенні довжини тіла на кожний сантиметр, ширина також збільшується в середньому на 0,7 см.

При графічному зображенні прямолінійних зв'язків виходить лінія, середня якої проходить по прямій.

При вимірі ступеня зв'язку між різними ознаками використовують їх нормовані відхилення, а коефіцієнт кореляції (r) має наступну просту формулу:

$$r = \frac{\sum_{j=1}^{n} x_{1j} \cdot x_{2j}}{\nu}$$

де $x_{ij} = \frac{V_{ij} - M}{\sigma_i}$ - j –те нормоване відхилення і-ї ознаки (i = 1, 2);

ν - число ступенів свободи $\nu = n - 1$; $\sigma_i = \sqrt{\frac{\sum_{j=1}^{n}(V_{ij} - M)^2}{\nu}}$

Сума добутків нормованих відхилень, що входить у формулу для коефіцієнта кореляції, володіє наступними трьома особливими властивостями:

1) якщо обидві ознаки змінюються паралельно, то сума добутків їх нормованих відхилень дає позитивну величину. Якщо при збільшенні однієї ознаки інший зменшується, то вся сума буде негативною.

Тому коефіцієнт кореляції визначає напрямок зв'язку: при прямих зв'язках він позитивний, а при зворотних - від'ємний;

2) при повних зв'язках, коли зміни обох ознак строго відповідають один одному і кореляційний зв'язок перетворюється на функціональний, сума добутків нормованих відхилень стає рівною кількості ступенів свободи: $\sum_{j=1}^{n} x_{1j} \cdot x_{2j} = \nu = n - 1$

Тому максимальне значення коефіцієнта кореляції дорівнює одиниці за абсолютною величиною $|r_{max}| = 1$:

для прямих зв'язків: r = +1;

для зворотних зв'язків: r = - 1;

3) при повній відсутності кореляційного зв'язку між ознаками коефіцієнт кореляції дорівнює нулю: $r_{min}=0$.

5.2 Помилка коефіцієнта кореляції

Як і всяка вибіркова величина, коефіцієнт кореляції має свою помилку репрезентативності, обчислювану для великих вибірок (n > 100) за формулою:

$$m_r = \frac{1 - (\bar{r})^2}{\sqrt{n-1}},$$

де $\bar{r}$ - коефіцієнт кореляції в генеральній сукупності, з якої взята вибірка; n - обсяг вибірки, тобто число пар значень, за якими обчислювався вибірковий коефіцієнт кореляції.

У більшості досліджень значення коефіцієнта кореляції в генеральній сукупності $\bar{r}$ невідомо, тому замість точного значення помилки коефіцієнта кореляції беруть наближене значення для вибіркового коефіцієнта кореляції r .

Приклад. При дослідженні 400 зерен кукурудзи знайдено, що коефіцієнт кореляції між довжиною і висотою зерна r = +0,85. Визначити, яка можлива величина коефіцієнта кореляції в генеральній сукупності.

Помилка знайденої величини:

$$m_r = \frac{1-0{,}85^2}{\sqrt{399}} = 0{,}014$$

Звідси при $t_1 = 2{,}0$ генеральний коефіцієнт кореляції:

$$\bar{r} = r \pm t \cdot m_r = 0{,}85 \pm 2 \cdot 0{,}014$$

Для малих вибірок (n < 100) необхідно користуватися іншою формулою помилки вибіркового коефіцієнта кореляції :

$$m_r = \sqrt{\frac{1-r^2}{n-2}}$$

Критерій достовірності коефіцієнта кореляції, що визначається за формулою:

$$t_r = \frac{r}{m_r} \geq t_{st}$$

оцінюється шляхом порівняння фактично отриманого значення t_r з фіксованими значеннями t_{st} (стандартне значення критерію Стьюдента), які відповідають трьом ступеням ймовірності безпомилкових прогнозів. При цьому для t_{st} ступінь свободи дорівнює ν = n - 2.

- якщо розрахункове $t_r > t_{st}$, то коефіцієнт кореляції достовірний, і можна вважати, що між досліджуваними ознаками існує взаємозв'язок;

- якщо $t_r < t_{st}$, то коефіцієнт кореляції недостовірний, і не можна зробити висновок про взаємозв'язок між досліджуваними ознаками у вибірці, а також у генеральній сукупності.

Приклад. Для з'ясування сили дії модифікуючих факторів (родючості грунту, кліматичних умов) при порівнянні двох сортів кукурудзи взяті 20 сусідніх ділянок, на яких попарно були висіяні один і другий порівнювані сорти, а потім розрахований коефіцієнт кореляції між врожаями порівнюваних сортів.

Великий коефіцієнт повинен вказувати на слабку дію модифікуючих агентів, а малий - повинен означати, що відмінності між парними ділянками по урожаю піддалися протягом досвіду якимось сильним і різноманітним впливам. У результаті були отримані наступні результати спостережень:

обсяг вибірки n = 20;

коефіцієнт кореляції r = +0,63.

Визначимо достовірність коефіцієнта кореляції:

$$m_r = \sqrt{\frac{1-0{,}63^2}{20-2}} = 0{,}18\,; \qquad t_r = \frac{0{,}63}{0{,}18} = 3{,}5$$

t_i={2,1 (b_1=0,95); 2,85 (b_2=0,99); 3,85 (b_3=0,999)}

Таким чином, отриманий достовірний коефіцієнт кореляції між врожаями сусідніх ділянок, для порога ймовірності не вище 0,999. Це означає, що відмінності дослідних ділянок за родючістю гранту та інших факторів, що визначає урожай, були слабкі і не дали проявитися відмінності випробовуваних сортів.

5.3 Приватний коефіцієнт кореляції

У деяких дослідженнях потрібно з'ясувати, чи не є зв'язок між двома ознаками обумовленої впливом якогї-небудь третьої ознаки. Наприклад, при вивченні статистичних зв'язків між урожаєм і середньою температурою повітря має сенс врахувати, вплив третьої ознаки - кількості опадів, що впливає на обидві ознаки - і на врожай, і на середню температуру повітря. Для того, щоб з'ясувати в таких дослідженнях, вплив третьої ознаки на кореляційний зв'язок між першим і другим ознакою, необхідно досліджувати цей зв'язок при його постійному значенні.

При постійному значенні ознаки можна тільки констатувати, що в мінливості інших ознак немає його впливу: він постійний, а інші ознаки змінюються.

Коефіцієнт кореляції між першим і другим ознаками при постійному значенні третьої ознаки називається приватним коефіцієнтом кореляції і позначається символом $r_{12\cdot3}$.

Для його розрахунку не завжди потрібно проводити розглянутий вище експеримент. Якщо зв'язок між парою ознак прямолінійний або відрізняється від прямолінійної незначно, то величину приватного коефіцієнта кореляції можна визначити за звичайними коефіцієнтами кореляції:

$$r_{12\cdot3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{\left(1 - r_{13}^2\right) \cdot \left(1 - r_{23}^2\right)}}, \text{ при } 3 = \text{const}$$

де: r_{12} - коефіцієнт кореляції між 1 і 2 ознакою;

r_{13} - коефіцієнт кореляції між 1 і 3 ознакою;

r_{23} - коефіцієнт кореляції між 2 і 3 ознакою.

Приклад. При дослідженні кореляційного зв'язку між вагою тварин (ознака 1) і діаметром мускульних волокон (ознака 2), без впливу на цей зв'язок калорійності їжі (ознака 3), (тобто при постійному значенні калорійності їжі) були отримані такі коефіцієнти кореляції: між вагою і діаметром волокон r_{12}=+0,6 (без вирівнювання калорійності їжі);

між вагою і калорійністю r_{13} = +0,8;

між діаметром волокон і калорійністю r_{23} = +0,7.

Приватний коефіцієнт кореляції:

$$r_{12\cdot3} = \frac{0{,}6 - 0{,}8 \cdot 0{,}7}{\sqrt{\left(1 - 0{,}8^2\right) \cdot \left(1 - 0{,}7^2\right)}} = 0{,}09; \qquad r_{13\cdot2} = 0{,}67 \qquad r_{23\cdot1} = 0{,}46$$

Виявилася дуже мала приватна кореляція. Дослідження показало, що якщо виключити статистичне вплив калорійності їжі, тобто вирівняти калорійність раціонів, то між вагою тварин і діаметром їх мускульних волокон не буде майже ніякої кореляції, хоча зазвичай, без вирівнювання калорійності їжі, цей зв'язок зовні виражається досить значним коефіцієнтом: +0,6.

При спільному вивченні трьох ознак можна виключити вплив не тільки третього, але також і першої або другої ознаки:

$$r_{13\cdot2} = \frac{r_{13} - r_{12} \cdot r_{23}}{\sqrt{\left(1 - r_{12}^2\right) \cdot \left(1 - r_{23}^2\right)}}, \text{ при } 2 = \text{const}$$

Іноді обчислення приватного коефіцієнта кореляції дає результати, що здаються на перший погляд неймовірними. Однак, при більш уважному аналізі явища, вже не з математичною, а зі спеціальною точки зору ці результати стають цілком зрозумілими і легко з'ясовними.

Приклад. При вивченні залежності ваги деревини (3) від розмірів дерева: обхвату (довжина периметра перетину) на рівні грудей (1) вимірює і висоти (2) стовбура - були отримані такі коефіцієнти кореляції:

- між обхватом (1) і висотою (2): r_{12}= +0,5;
- між обхватом (1) і вагою (3): r_{13}= + 0,9;

між висотою (2) і вагою (3): r_{23}= + 0,8.

Окремі коефіцієнти кореляції кожного розміру з вагою при виключеному впливі іншого розміру не викликають ніяких здивувань і вказують на велику приватну кореляцію обхвату і висоти з вагою деревини:

$$r_{13\cdot2} = \frac{0{,}90 - 0{,}50 \cdot 0{,}80}{\sqrt{(1-0{,}25)\cdot(1-0{,}64)}} = +0{,}96$$

$$r_{23\cdot1} = \frac{0{,}80 - 0{,}50 \cdot 0{,}90}{\sqrt{(1-0{,}25)\cdot(1-0{,}81)}} = +0{,}92$$

Приватна кореляція між обома розмірами при виключеному впливі ваги, тобто при її постійному значенні:

$$r_{12\cdot3} = \frac{0{,}50 - 0{,}90 \cdot 0{,}80}{\sqrt{(1-0{,}81)\cdot(1-0{,}64)}} = -0{,}84$$

Виявилося, що між обхватом і висотою дерева вийшла значна негативна приватна залежність: при збільшенні висоти, обхват дерева зменшується. Це, здавалося б, явно суперечить звичайним процесам розвитку дерев: якщо збільшується висота, то, звичайно, збільшується і обхват.

Пояснення цього уявного протиріччя полягає в основній умові приватної кореляції - сталості виключаємої ознаки.

Якщо взяти дерева одного і того ж ваги, то серед таких дерев збільшення висоти може відбуватися тільки за рахунок зменшення обхвату. Якби збільшувалися обидва розміри, то вага деревини не могла би залишатися постійною.

При кореляції 4 -х ознак, розрахунок ведуть за формулою:

$$r_{12\cdot34} = \frac{r_{12\cdot4} - r_{13\cdot4} \cdot r_{23\cdot4}}{\sqrt{\left(1-r_{13\cdot4}^2\right)\cdot\left(1-r_{23\cdot4}^2\right)}}$$

Множинний коефіцієнт кореляції показує взаємозв'язок між усіма досліджуваними ознаками одночасно.

5.4 Помилка приватного коефіцієнта кореляції

Помилка репрезентативності вибіркового приватного коефіцієнта кореляції розраховується за такою ж формулою, як і у випадку звичайного коефіцієнта кореляції для нечисленні груп (n < 100):

$$m_{r_{12}} = \sqrt{\frac{1-r^2}{n-2}}$$

При оцінці критерію достовірності приватного коефіцієнта кореляції, граничні значення показника ймовірності беруться для числа ступенів свободи, які відповідають ν = n - 2 - k, де k - число елімінуваних ознак.

Обсяг вибірки в даному випадку дорівнює числу пар значень (n), однаковому для всіх звичайних коефіцієнтів кореляції, які необхідні для розрахунку приватного коефіцієнта кореляції.

5.5 Коефіцієнт прямолінійною регресії

Прямолінійна кореляція відрізняється тим, що при цій формі зв'язку кожному з однакових змін першої ознаки відповідає цілком визначене і теж однакова в середньому зміна іншої ознаки, пов'язаної з першою або залежної від першої.

Та величина, на яку в середньому змінюється друга ознака, при зміні першої на одиницю виміру, називається коефіцієнтом прямолінійною регресії. Розраховується вона за такою формулою:

$$R = \frac{\sigma_2}{\sigma_1} \cdot r_{12},$$

де R - коефіцієнт прямолінійної регресії;

σ_2 - середньоквадратичне відхилення другої ознаки, яке змінюється у зв'язку зі зміною першого;

σ_1 - середньоквадратичне відхилення першої ознаки, у зв'язку зі зміною якої змінюється друга ознака;

r_{12} - коефіцієнт кореляції між першою і другою ознаками.

Помилка коефіцієнта регресії дорівнює помилці коефіцієнта кореляції, помноженої на відношення сигм:

$$m_R = \frac{\sigma_2}{\sigma_1} \cdot m_r = \frac{\sigma_2}{\sigma_1} \cdot \sqrt{\frac{1-r^2}{n-2}}$$

Критерій достовірності коефіцієнта регресії дорівнює критерію достовірності коефіцієнта кореляції:

$$t_R = \frac{R}{m_R} = \frac{\frac{\sigma_2}{\sigma_1} \cdot r_{12}}{\frac{\sigma_2}{\sigma_1} \cdot m_r} = \frac{r}{m_r} = t_r$$

Приклад. При розробці методів селекції молочної худоби з'ясовувався зв'язок вищого добового надою з удоєм за 300 днів тієї ж лактації. Всього вивчено 577 лактацій, що проходили в оптимальних умовах. Були отримані наступні дані:

- вищий добовий удій (ознака 1): $n_1 = 57$; $M_1 = 17{,}2$ кг; $\sigma_1 = 3{,}9$ кг;
- удій за 300 днів лактації (ознака 2): $n_2 = 577$; $M_2 = 3250$ кг; $\sigma_2 = 685$ кг;
- коефіцієнт кореляції: $r_{12} = +0{,}829$.

Розрахуємо коефіцієнт регресії удою за 300 днів по вищому добовому удою:

$$\widetilde{R} = \frac{685}{3,9} \cdot (+0,829) = +145,6\ кг \qquad m_R = \frac{685}{3,9} \cdot \sqrt{\frac{1 \cdot 0,829^2}{575}} = 4,2$$

$$\overline{R} = \widetilde{R} \pm 2m_R = +145,6 \pm 2 \cdot 4,2 \quad \begin{cases} \text{не менее} + 137,2\ кг \\ \text{не более} + 153,8\ кг \end{cases}$$

Обчислення показали, що в даному випадку генеральний коефіцієнт регресії дорівнює R = +145,6 ± 4,2 кг. Це означає, що при збільшенні вищого добового надою на кожен 1 кг удій за 300 днів лактації збільшується на +145,6 кг з можливими відхиленнями цієї величини в межах 138 - 154 кг.

Таким чином, якщо, наприклад, у групи корів вищий добовий удій в середньому на 5 кг більше середнього по одноліток, то можна очікувати, що удій за 300 днів лактації цих корів буде на 5 • 145,6 = 728 кг більше середнього по їх одноліток.

5.6 Тетрахорічний показник зв'язку

При альтернативному розмаїтті, коли обидві якісні ознаки виражаються тільки наявністю або відсутністю їх у особин, кореляційний зв'язок між двома ознаками вимірюється тетрахорічним показником зв'язку.

Якщо у кожної особини вивчаються дві ознаки, то вся група розбивається на наступні чотири частини:

[a] – особини, які мають обидві ознаки (+ +);

[b] – особини, які мають першу ознаку, але не мають другої (+ -);

[c] – особини, що не мають першої ознаки, але мають другу (- +);

[d] - особини, які не мають обох ознак (- -).

Якщо позначити чисельність зазначених чотирьох груп цими ж буквами (a, b, c, d), то ступінь зв'язку наявності першої ознаки з наявністю другої ознаки визначатиметься тетрахорічним показником зв'язку, який обчислюється за такою формулою:

$$r_{++} = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (a+c) \cdot (d+b) \cdot (d+c)}}$$

Приклад. При перевірці ефективності дії щеплення проти висипного тифу отримані первинні матеріали про кількість хворих (-) і не хворих (+) з числа хто отримували (+) і не отримували (-) щеплення (див. табл.5.1). Обсяг вибірки n = 210 осіб.

Таблиця 5.1

	Признак 1		Σ
Признак 2	Отримали щеплення (+)	Не отримали щеплення (-)	
Не захворіли (+)	(a)++=54	(c)-+=106	(a+c)=160
Захворіли (-)	(b)+-=6	(d)--=44	(b+d)=50
Σ	(a+b)=60	(c+d)=150	N=210

$$r_{++} = \frac{54 \cdot 44 - 6 \cdot 106}{\sqrt{60 \cdot 150 \cdot 160 \cdot 50}} = \frac{+1740}{8485,3} = +0,205$$

Достовірність тетрахорічного коефіцієнта кореляції визначається за величиною χ^2_{r++} (ксі).

$$\chi^2_{r++} = n \cdot r^2_{++}, \qquad \chi^2_{r++} \geq \chi^2_{st}$$

n - загальний обсяг вибірки.

Наприклад, припустимо n = 210, тоді:

$\chi^2_{r++}=210 \cdot 0{,}205^2=8{,}8$

Величині χ^2_{r++} має відповідати табличне значення (табл. 5.2), яке залежить від відповідальності порога ймовірності (β). Число ступенів свободи для тетрахорічного коефіцієнта дорівнює 1, тобто число градацій мінус одиниця.

Таблиця 5.2 – Стандартне значення χ^2_{st}

β	0,95	0,99	0,999
χ^2_{st}	3,8	6,6	10,8

5.7 Поліхорічний показник зв'язку

Існують такі кількісні ознаки, ступінь розвитку яких характеризується не результатом точного вимірювання, не числом, а якісними градаціями, які визначаються суб'єктивно, шляхом огляду або смакової проби.

Наприклад:

а) колір пера птиці - світло-сірий , сірий і темно-сірий;

б) смак вершкового масла - слабо- , середньо- і сильносолений;

в) вгодованість тварин - жирна, вищесередньої, середня, нижчого за середній, худа і т.д.

Визначення ступеня кореляційної зв'язку між такими ознаками можна проводити за допомогою поліхорічного показника зв'язку, що позначається грецькою буквою ρ і обчислюваного за такою формулою:

$$\rho = \frac{a-1}{\sqrt{(gr_1-1)\cdot(gr_2-1)}}, \qquad \text{где:} \quad a = \sum\left(\frac{\sum \frac{f^2}{n_2}}{n_1}\right);$$

де f - частоти осередків кореляційної решітки по першій і другій ознакам;

n_1 - частоти ряду першої ознаки (визначаються за стовпцями за нижньому сумарному рядку кореляційної таблиці-решітки);

n_2 - частоти ряду другої ознаки (визначаються по рядках в правому сумарному стовпці кореляційної таблиці-решітки);

gr_1, gr_2 - число градацій, на які розбиті перший і другий ознаки;

n - загальна чисельність групи: $n=\Sigma n_1=\Sigma n_2$.

Поліхорічний показник зв'язку завжди виражається позитивним числом, тому визначення характеру зв'язку (пряма/зворотна) проводиться по виду кореляційної решітки.

Приклад. При дослідженні зв'язку між міцністю статури одного виду тварин (ознака 1) і густотою їх шерсті (ознака 2) отримані наступні дані (див. табл. 5.3).

Таблиця 5.3 – Залежність ознаки 1 від ознаки 2

Шерсть (ознака 2)	Міцність статури (ознака 1)			n_2
	Міцна	Середня	Слабка	
Густа	30	9	1	40
Середня	5	21	4	30
Рідка	2	3	25	30
n_1	37	33	30	100

Розрахунок поліхорічного показника зв'язку потрібно виробляти за такими етапами:

1. Підрахувати частоти n_1 за першою ознакою - суми по стовпцях (37, 33, 30), потім частоти n_1 по другому - суми по рядках і загальну чисельність групи $n = n_1 = n_2 = 37+33+30 = 40+30+30 =100$.

2. У кожній клітинці звести в квадрат частоту і отриманий результат f^2 записати в тому ж осередку в дужках. Потім квадрат частоти осередку розділити на частоту другої ознаки по тому ж рядку, в якій знаходиться осередок, і отриманий результат $\frac{f^2}{n_2}$ записати в тій же комірці під раніше записаної цифрою (див. табл. 5.4).

3. Останні числа осередків $\frac{f^2}{n_2}$ скласти по стовпцях, тобто по градаціях другої ознаки.

4. Отримані значення розділити на частоти ряду першої ознаки n_1.

5. Знайти суму значень цифр останнього рядка. Це буде величина *a*.

6. Значення *a*, gr_1, gr_2, n підставити у формулу для поліхорічного показника зв'язку.

Таблиця 5.4 – Результати розрахунку

Шерсть (ознака 1)	Міцність статури (ознака 2)			n_2
	Сильна	Середня	Слабка	
Густа	f^2=(900) f^2/n_2=22,5	(81) 2,02	(1) 0,02	40
Середня	(25) 0,83	(441) 14,7	(16) 0,53	30
Рідка	(4) 0,13	(9) 0,30	(625) 20,83	30
n_1	37	33	30	100
$\Sigma(f^2:n_2)$	23,46	17,02	21,38	-
$\frac{\Sigma(f^2:n_2)}{n_1}$	0,63	0,52	0,71	a=1,86

Зведення матеріалів в кореляційну решітку виявило цілком помітний зв'язок між досліджуваними ознаками: при сильній міцності статури більшість особин мало густу шерсть. Ступінь зв'язку між цими ознаками визначимо, розрахувавши поліхорічний показник зв'язку:

$$gr_1 = 3; \qquad gr_2 = 3; \qquad n=100; \qquad \rho = \frac{1{,}86-1}{\sqrt{(3-1)\cdot(3-1)}} = 0{,}43$$

Між вивченими ознаками мається кореляційний зв'язок = 0,43.

Достовірність поліхорічного показника зв'язку можна визначити за допомогою критерію χ^2, який для даного показника дорівнює $\chi = n \cdot (a-1) \geq \chi^2_{st}$ при числі ступенів свободи $\nu=(gr_1-1)\cdot(gr_2-1)$. Для нашого прикладу:

$$\chi^2 = 100 \cdot (1{,}86 - 1) = 86{,}0 \qquad \chi^2_{st} = (18{,}5 \;\; 13{,}3 \;\; 9{,}5) \;\; \text{для}\; \nu = 2 \cdot 2 = 4$$

$$\chi^2 > \chi^2_{st}$$

5.8 Перевірка артефактів (випадів)

Артефактом (випадом) - називається значення ознаки, що різко виділяється (дуже велике або маленьке) на тлі інших значень цієї ознаки.

Перевірка артефактів повинна бути проведена перед початком обробки будь-яких експериментальних даних. Якщо підтвердиться, що значення сильно випадають (виділяються) не можуть ставитися до об'єктів даної групи і потрапили в записі спостережень за помилок уваги, то такий артефакт виключається з обробки. Перевірка на артефакт проводиться виходячи з наступного умови:

$$T = \frac{\left|\hat{V} - M\right|}{\sigma} \geq T_{st},$$

T - критерій артефакту; $\hat{V}$ - значення ознаки, що виділяється (можливий артефакт); T_{st} - табличне значення критерію артефакту; σ - середньоквадратичне відхилення (без артефакту); M - середнє значення (без артефакту).

Стандартні значення критерію випадів зведені в табл. 5.5.

Таблиця 5.5 - Стандартні значення критерію випадів

N	2	3-4	4-9	10-15	16-20	21-28	29-34	35-46
T_{st}	2	2,1	2,2	2,3	2,4	2,5	2,6	2,7
N	47-66	67-84	85-104	105-124	125-174	175-349	350-599	600-1500
T_{st}	2,8	2,9	3	3,1	3,2	3,3	3,4	3,5

Приклад. Отримано значення ознаки: 1, 2, 3, 10. Визначити чи є число 10 у цій послідовності артефактом.

n=3, M=3,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(V_i - M)^2}{n-1}} = \sqrt{\frac{(1-2)^2 + (2+2)^2 + (3-2)^2}{3-1}} = 1$$

$$T = \frac{10-2}{1} = 8 \geq 2{,}1$$

Таким чином, число 10 випадає і його можна виключити з розгляду.

6 ДИСПЕРСІЙНИЙ АНАЛІЗ

Для того, щоб, користуючись дисперсійним аналізом, отримати правильні результати, необхідно виконувати певні правила організації дисперсійних комплексів.

Дисперсійний аналіз полягає у вивченні статистичного впливу одного або декількох факторів на результативну ознаку (у).

Результативна ознака (у) - це ознака, яка вивчається як результат статистичного впливу факторів: організованих (контрольованих або облікованих) (х) і всіх інших неорганізованих в даному дослідженні (z).

Результативними ознаками можуть бути:

- точно вимірювані кількісні особливості об'єктів - довжина, ширина, жвавість, шерстность;
- неточно вимірювані особливості - густота крему, колір, розумові здібності;
- комбіновані ознаки - співвідношення розмірів тіла, індекси продуктивності;
- якісні ознаки - масть, хвороба, одужання.
- окремі ознаки, що приймаються за аргумент при вивченні середнього ознаки, що приймається за функцію.

Фактор - це будь-який вплив, вплив або стан, різноманітність яких може відбиватися на різноманітності результативної ознаки.

Фактори можуть бути:

- фізичного впливу - температура, вологість, тиск;
- хімічного впливу - харчування
- біологічного впливу - наявність мутагенів, вік, стать, сорт, національність.

Градаціями факторів називається ступінь їх дії чи стану об'єктів вивчення. В якості градації факторів можуть виступати різна температура, вологість, доза опромінення, різна тривалість впливу, різна поживність і склад корму, дози стимулюючих і хімічних мутагенів, різні періоди хвороби, ступінь таланту, різні батьки, різні ареали проживання, різні умови життя.

Градаціями комплексу називаються опитні групи досліджень. Кожна градація комплексу відповідає одній градації фактора і включає ті об'єкти з їх датами, які піддаються одній ступені впливу фактора або знаходиться в одному з досліджуваних станів. Організація градацій комплексу здійснюється різними способами, такими як підбір опитних і контрольних груп, використання раніше отриманих результатів дослідження, систематизація записів виробничої звітності.

Зазвичай комплекс представляють у вигляді таблиці, стовпці якої відповідають градаціях (комбінаціям) чинників.

Різноманітність - це наявність неоднакових значень кожної ознаки у різних особин об'єднаних в одну групу. Як зазначалося раніше, різноманітність групи особин по досліджуваній ознаці вимірюється показниками різноманітності: лімітами, середнім квадратичним відхиленням, коефіцієнтами варіації.

Для того, щоб з'ясувати ступінь і достовірність впливу досліджуваних організованих і неорганізованих факторів, вимірюють ту частину загальної різноманітності, яка викликається цими факторами. Робиться це за допомогою двох величин: дисперсії і девіати.

Дисперсія - це первинна міра різноманітності в розглянутій групі. Вона дорівнює сумі квадратів центральних відхилень. Загальна дисперсія ознаки визначається як:

$$C_y = \sum_{i=1}^{n} (V_i - M)^2$$

Загальна різноманітність результативної ознаки завжди більше того розмаїття, яке пов'язане зі статистичним впливом організованих факторів. Відбувається це тому, що в будь-якому дослідженні не можна звільнитися від дії всієї безлічі інших чинників, так чи інакше впливаючих на зміну результативної ознаки.

Тому при проведенні дисперсійного аналізу загальна дисперсія ознаки Су у досліджуваній групі розчленовується на дві дисперсії - факторіальна або приватну (викликану організованими факторами) C_x, і випадкову або залишкову дисперсію (викликану іншими, неорганізованими в даному досвіді факторами) C_z. Сума факторіальної і випадкової дисперсій завжди дорівнює загальній:

$$C_y=C_x+C_z$$

Факторіальна дисперсія :

$$C_x = \sum (M_x - M)^2$$

Випадкова дисперсія :

$$C_z = \sum (V - M_x)^2$$

де M_x - приватна середня результативної ознаки по кожній окремій градації організованих факторів.

Дисперсія як показник різноманітності залежить від числа особин в групі. Для визначення ступеня впливу факторів ця обставина не має значення. Для інших же цілей, зокрема для встановлення достовірності впливу чинників, виявленого при вибірковому дослідженні, необхідний показник, вільний від зазначеної залежності, що допускає порівняння груп, різних за кількістю вхідних в них елементів. Таким показником є девіата.

Девіатой називають дисперсію, що припадає на один елемент вільного різноманіття або на одну ступінь свободи:

$$\sigma_y^2 = \frac{C_y}{\nu_y} \qquad \sigma_x^2 = \frac{C_x}{\nu_x} \qquad \sigma_z^2 = \frac{C_z}{\nu_z}$$

де σ_y^2 - загальна девіата по всьому комплексу; σ_x^2 - факторіальна девіата; σ_y^2 - випадкова девіата; ν - число ступенів свободи.

Корінь квадратний з девіати є середнім квадратичним відхиленням:

$$\sigma = \sqrt{\sigma_y^2}$$

Девіати використовуються в дисперсійному аналізі для визначення достовірності впливу організованих факторів на результативну ознаку, виявлену у вибірковому дослідженні. Достовірність впливу організованого фактора визначається відношенням факториальной девіати до випадкової, яка повинна бути не менше табличній стандартної величини F_{st}:

$$F = \frac{\sigma_x^2}{\sigma_z^2} \geq F_{st}$$

Якщо це відношення дорівнює або більше визначеної стандартної величини F_{st}, вплив вважається достовірним з певним ступенем імовірності. Стандартні відносини девіат F_{st} визначаються за спеціальними таблицями .

Приклад. Вивчається дія на ріст рослин (у), який є результативною ознакою двох факторів: А (температура середовища) і В (вологість). Кожен фактор береться у двох градаціях: A_1 і A_2 - низька і висока температура; B_1 і B_2 мала і велика вологість.

Для кожної з чотирьох градацій двох факторів - A_1B_1, A_1B_2, A_2B_1, A_2B_2 - за способом випадкової вибірки вибрано по дві особини. У всіх восьми особин виміряно результативну ознаку і результати записані у вигляді статистичного комплексу в табл.6.1.

Таблиця 6.1 – Результати вимірів

Градація 1-го фактора	A_1				A_2			
Градації 2-го фактора	B_1		B_2		B_1		B_2	
Значення результативної ознаки	9	11	3	5	1	3	7	9

Необхідно проаналізувати отриманий комплекс і встановити наступне: чи справляють вплив на результативну ознаку - зростання, чинники, що вивчаються - температура і вологість в їх загальної сумарної дії, яка роль кожного фактора окремо і в їх поєднаннях?

Для вирішення, спочатку припустимо, що діють не два, а один сумарний фактор *x*, що має всі зазначені 4-ре градації, які організовані у дослідженні для всіх розглянутих факторів.

Знайдемо загальну дисперсію C_y. Для цього розрахуємо загальну середню:

$$M = \frac{9+11+3+5+1+3+7+9}{8} = 6$$

тоді: $C_y=(9-6)^2+(11-6)^2+(3-6)^2+(5-6)^2+(1-6)^2+(3-6)^2+(7-6)^2+(9-6)^2=88$

Знайдемо факторіальна дисперсію. Розрахунок зведемо в табл. 6.2.

Таблиця 6.2 – Результати розрахунку

Градації факторів	x_1		x_2		x_3		x_4		Підсумок
	x_{11}	x_{12}	x_{21}	x_{22}	x_{31}	x_{32}	x_{41}	x_{42}	
Значення результативної ознаки	9	11	3	5	1	3	7	9	n=8
Сума значень	20		8		4		16		Σ=48
Приватні середні M_x	$M_{x1}=10$		$M_{x2}=4$		$M_{x3}=2$		$M_{x4}=8$		M=6
Центральні відхилення середніх (M_x - M)	+4		-2		-4		+2		
$(M_x - M)^2$	16	16	4	4	16	16	4	4	$C_x=80$
Приватні центральні відхилення (V- M_x)	-1	+1	-1	+1	-1	+1	-1	+1	
$(V- M_x)^2$	1	1	1	1	1	1	1	1	$C_z=8$

Кількість квадратів центральних відхилень дат беруть стільки, скільки дат результативної ознаки (= 8). На рис. 6.1 представлена залежність змінення приватної середньої від градації факторів.

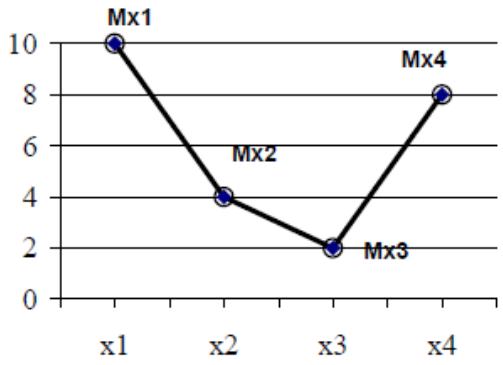


Рисунок 6.1 – Змінення приватної середньої від градації факторів

Ступінь впливу фактора на результативну ознаку дорівнює відношенню факториальной дисперсії до загальної:

$$\eta_x^2 = \frac{C_x}{C_y} = \frac{80}{88} = 0{,}91$$

Отриманий результат означає, що 91 % всього різноманіття результативної ознаки визначається різноманітністю організованих факторів.

Знайдемо випадкову дисперсію. Розрахунок зведений в табл. 6.2. Ступінь впливу неорганізованих факторів:

$$\eta_x^2 = \frac{C_z}{C_y} = \frac{8}{88} = 0{,}09$$

Таким чином, вплив неорганізованих факторів становить всього 9 % від загального впливу всіх факторів. Це вказує на велику силу впливу сумарного фактора у вигляді температури і вологості на результативну ознаку - зростання.

Графік показує, що при зміні фактора від x_1 до x_4 результативна ознака спочатку зменшується, а потім зростає. Такий вплив одного не об'єднаного фактора зустрічається рідко, але в нашому випадку градаціями факторами x є чотири комбінації градацій двох факторів.

Визначимо достовірність впливу організованого фактора за допомогою девіати.

а) для загальної дисперсії число ступенів свободи дорівнює:

$$\nu = n-1=8-1=7,$$

де n - число градацій результативної ознаки (y);

б) для факториальної дисперсії число ступенів свободи дорівнює:

$$\nu_x = r_x-1=4-1=3,$$

де r_x - число градацій організованого сумарного фактора;

факторіальна девіата:

$$\sigma_x^2 = \frac{C_x}{\nu_x} = \frac{80}{3} = 26,7$$

в) для випадкової дисперсії число ступенів свободи одно:

$$\nu_z = n - r_x = 8\text{-}4 = 4$$

випадкова девіата:

$$\sigma_x^2 = \frac{C_z}{\nu_z} = \frac{8}{2} = 2,0$$

Достовірність впливу: $F = \frac{\sigma_x^2}{\sigma_z^2} = \frac{26,7}{2} = 13,4 \geq F_{st} = 6,6(b_1 = 0,95)$

для $b_2 = 0,99$ - $F_{st} = 16,7$
для $b_3 = 0,999$ - $F_{st} = 56,1$

Це означає, що спостережуємий вплив організованого сумарного фактора з достовірністю $b_1 = 0,95$ не є випадковим, так як факторіальна дисперсія виявилася більшою ніж випадкова.

Таким чином, сумарна дія двох факторів А і В на результативну ознаку дуже велике і достовірно. Залишається з'ясувати:

1) яке значення кожного з цих факторів окремо при вирівнянній дії іншого;
2) яке значення відмінностей їх спільної дії при різних комбінаціях градацій.

Для пошуку відповідей на ці питання використовують розгорнутий комплекс, який називається двухфакторним.

Рішення двухфакторного комплексу на розглянутому прикладі проводиться за такими етапами:

1. виконують визначення загальної дисперсії - аналогічно однофакторном комплексу:

$$C_y = 88$$

2. виконують визначення випадкової дисперсії C_z - аналогічно одно факторному комплексу:

- приватні середні M_x по 4 - ма градаціях: 10, 4, 2, 8;
- відхилення від своєї приватної середньої $(V - M_x)$: -1, +1, -1, +1, -1, +1, -1, +1;
- квадрати відхилень $(V - M_x)^2$: +1, +1, +1, +1, +1, +1, +1, +1;
- їх сума $\Sigma(V - M_x)^2$: $C_z = 8$.

3. Визначення дисперсії сумарної дії організованих факторів аналогічно однофакторному комплексу (див. табл. 6.2): $C_x = 80$.

4. Визначення приватних факторіальних дисперсій окремо по кожному фактору:

- розрахунок приватних середніх M_A і дисперсій за фактором А:

$$C_A = \Sigma(M_A - M)$$

- розрахунок приватних середніх M_B і дисперсій за фактором В:

$$C_B = \Sigma(M_B - M)$$

Результати зводимо у табл. 6.3.

Таблиця 6.3 - Дисперсійний аналіз двухфакторного комплексу

n=8	Градація 1-го фактору	A_1				A_2			
ΣV=48	Градація 2-го фактору	B_1		B_2		B_1		B_2	
	Значення результативної ознаки (зростання)	9	11	3	5	1	3	7	9
M=6	M_x	$M_{x1}=10$		$M_{x2}=4$		$M_{x3}=2$		$M_{x4}=8$	
	(M_x-M)	+4		-2		-4		+2	
$C_x=80$	$(M_x-M)^2$	16	16	4	4	16	16	4	4
Градації фактора А		A_1				A_2			
	Значення по градаціях	9	11	3	5	1	3	7	9
	Сума по градаціях А	28				20			
	Приватні середні M_A	$\frac{28}{4}=7$				$\frac{20}{4}=5$			
	(M_A-M)	7-6=+1				5-6=-1			
	$(M_A-M)^2$	1	1	1	1	1	1	1	1
Градації фактора В		B_1				B_2			
	Значення по градаціях	9	11	1	3	3	5	7	9
	Сума по градаціях В	24				24			
	Приватні середні M_B	$\frac{24}{4}=6$				$\frac{24}{4}=6$			
M=6	(M_B-M)	6-6=0				6-6=0			
$C_B=0$	$(M_A-M)^2$	0	0	0	0	0	0	0	0

5. Ступінь впливу кожного фактора:

$$\eta_A^2=\frac{C_A}{C_y}=\frac{8}{88}=0{,}09 \qquad \eta_B^2=\frac{C_B}{C_y}=\frac{0}{88}=0$$

Отриманий результат свідчить про те, що вплив фактора В при вирівняних значеннях фактора А не виявляється в різноманітності результативної ознаки.

Таке співвідношення показників η_A^2 = 9 % , η_B^2 = 0 %, η_x^2 = 91 % відображає вплив одних факторів на інші і показує, що при нормальній температурі A_1 нормальна вологість B_1 сприятлива для зростання, а підвищена вологість B_2 вже пригнічує ріст, тобто при поєднанні градацій A_1B_1:M_x = 10, а при поєднанні A_1B_2:Mx = 4.

При підвищеній температурі A_2, навпаки, низька вологість B_1 недостатня для нормального росту (поєднання A_2B_1:M_x = 2), і за цим він уповільнений, а підвищена вологість B_2 сприятлива, і зростання посилюється (поєднання A_2B_1:M_x = 8).

Якщо розглядати кожен чинник окремо, то температура без регулювання вологості і вологість без регулювання температури самі по собі слабко виявляються в різноманітності зростання: η_A^2 = 9 % , η_B^2 = 0 %.

Якщо організувати певне поєднання чинників (їх градацій), наприклад, при певній температурі забезпечити певну вологість, то різні комбінації цих факторів створять значну різноманітність результативної ознаки (росту рослин), що і покаже їх велика сумарна дія: η_x^2 = 91 %. Оскільки завжди є деяка відмінність у дії одного фактора при різних градаціях іншого, то в кожної градації дисперсійного комплексу сумарна дія всіх організованих факторів складається з дії кожного фактора окремо і специфічної дії від їх поєднань.

6.1 Підбір факторів для дисперсійного аналізу

При організації однофакторних комплексів фактором вважається будь-яка ознака, вплив якої на результативну ознаку потрібно вивчити. Це можуть бути інші ознаки тієї ж тварини або рослини, різні умови життя, хімічні або біологічні агенти та інші впливи.

При організації двох - і багатофакторних комплексів вільний вибір факторів для дослідження обмежений вимогою повної незалежності їх між собою. Для таких комплексів не можна в якості двох факторів брати, наприклад, вагу і розмір тварин, так як ці ознаки не можна підбирати незалежно один від одного: при малій вазі неможливо підібрати такі ж значення розміру, як і при великій вазі.

Незалежними факторами можуть бути, наприклад, температура а вологість, стать і рівень годівлі, хімічне та біологічне вплив.

6.2 Поділ факторів на градації

При проведенні дисперсійного аналізу не потрібно, щоб фактори були розділені обов'язково на кількісні градації у формі варіаційного ряду. Як для однофакторних, так і для двох- і багатофакторних комплексів фактори можуть мати і якісні градації, наприклад, стать - чоловіча, жіноча; колір волосся або пера світло-сірий, сірий, темно-сірий; вгодованість - жирна, вище середньої, середня, нижче середньої; міцність статури - слабка, нормальна, сильна.

При встановленні градації факторів потрібно пам'ятати, що результати дисперсійного аналізу у великій мірі залежать від того рівня, на якому встановлені градації факторів.

Якщо, наприклад вивчається дія температури - то при градаціях 15 °, 20 °, 25 °С може бути знайдено достовірний вплив цього фактора на результативну ознаку, але це зовсім не означає, що такий ж сильний вплив буде при іншому рівні градацій, наприклад, 5 °, 10 °, 15 °С.

Велике значення також має рівень групи неорганізованих факторів, які складають фон дисперсійного аналізу. Наприклад, комбінована дія віку і якого-небудь стимулятора ожиріння дають при одному рівні загальної годівлі та утримання певний ефект, а при іншому, наприклад мізерному годуванні і поганому змісті, може і зовсім не проявитися.

6.3 Підбір особин. Типи комплексів

Результати дисперсійного аналізу в основному залежать від того, наскільки правильно підібрані особини, як за якістю, так і за кількістю. За своєю якістю особини для дисперсійного аналізу повинні відображати ту генеральну сукупність, для вивчення якої і проводиться дослідження.

За величиною результативного знаки особини повинні бути підібрані за принципом випадкової вибірки. Найкраще при відборі об'єктів для дисперсійного аналізу поступати таким чином.

Нехай для даної градації потрібно, наприклад, 20 особин, а всього мається 30 особин. Тоді номер кожної особини потрібно записати на картку. Всі 30 карток добре перетасувати і взяти поспіль без вибору перші 20 карток або, навпаки, взяти поспіль тільки перші 10 карток. У першому випадку відберуть особини для дослідження, у другому відкинуться особини, зайві для даної градації.

Організація дисперсійного комплексу з виконанням принципу випадковості відбору варіантів називається рандомизацией, а комплекси, організовані таким чином, називаються рандомізованими.

За кількістю особини можуть розподілятися по градаціях факторів різними способами: порівну, пропорційно, нерівномірно. Відповідно до цього організовані

комплекси бувають рівномірними (підбирається однакове число дат) і нерівномірними (підбирається неоднакове число дат).

Якщо градації двох або багатофакторних комплексів заповнені різним числом дат, але таким чином, що дати по градації одного фактора знаходились в однаковому відношенні (пропорції) до всієї решти чинників, то комплекс називається пропорційним (табл. 6.4). Рівномірні і пропорційні комплекси називаються ортогональними. Рівномірний комплекс є приватним випадком пропорційного, коли ставлення частот дорівнює 1:1; 1:1; 1:1 і т.д. (табл. 6.5).

Таблиця 6.4 – Пропорціональний комплекс

A_1		A_2	
B_1	B_2	B_1	B_2
2	4	6	12
1 : 2		1 : 2	

Таблиця 6.5 – Рівномірний комплекс

A_1		A_2	
B_1	B_2	B_1	B_2
2	2	8	8
1 : 1		1 : 1	

Для рівномірних і пропорційних комплексів сума приватних дисперсій дорівнює загальній:

$$C_A+C_B+C_{AB}=C_x$$

У деяких випадках буває легко організувати пропорційний комплекс на основі наявного нерівномірного. Розглянемо, як це зробити на наступному прикладі.

Припустимо, що є деяка кількість особин, які за своєю якістю відповідають вимогам градацій двухфакторного комплексу, але за кількістю не відповідають вимогам пропорційності (див. табл.6.6).

Таблиця 6.6 – Вихідні данні

	A_1(1)			A_2 (3)			Сума частот по градаціям	Число пропорційності
	B_1	B_2	B_3	B_1	B_2	B_3	3+10+13=26 (A_1)	$k_{A1}=\frac{26}{26}=1$
n_x	3	10	13	11	30	39		
	$\frac{3}{3}\div\frac{10}{3}\div\frac{13}{3}=$ $=1\div3{,}3\div4{,}3$			$\frac{11}{11}\div\frac{30}{11}\div\frac{39}{11}=$ $=1\div2{,}7\div3{,}6$			11+30+39=80 (A_2)	$k_{A2}=\frac{80}{26}=3{,}1\approx3$
							3+11=14 (B_1)	$k_{B1}=\frac{14}{14}=1$
							10+30=40 (B_2)	$k_{B2}=\frac{40}{14}=2{,}9\approx3$
							13+39=52 (B3)	$k_{B3}=\frac{52}{14}=3{,}7\approx4$

Тут ставлення частот фактора В за різними градаціях фактора А неоднаково. Для A_1 відношення частот дорівнює 1:3,3:4,3; для A_2 - 1: 2,7: 3,6. Але в цьому нерівномірному комплексі ставлення частот по градаціях кожного фактора окремо (що показано в правій частині табл.6.6) близькі до певних цілим числам:

$A_1:A_2 \approx 1:3$ $B_1:B_2:B_3 \approx 1:3:4$

Ці числа пропорційності (k_{Ai}, k_{Bi}) можна використовувати для побудови пропорційного комплексу.

Для цього в кожній градації за обома факторами треба перемножити відповідні числа пропорційності, потім фактичні частоти розділити на ці добутки і взяти найменше з отриманих приватних (a_{min}). Цю величину треба помножити на добутки чисел пропорційності. Таким чином, отримують частоти пропорційного комплексу, який утворений з наявного непропорційного з найменшою вибраковуванням особин. Ці дії показані в наступній табл. 6.7.

Таблиця 6.7 - Організація пропорційного дисперсійного комплексу на основі наявного непропорційного

Градація 1-го фактору	$A_1(1)$			$A_2(3)$			
Градація 2-го фактору	$B_1(1)$	$B_2(3)$	$B_3(4)$	$B_1(1)$	$B_2(3)$	$B_3(4)$	
Фактичні частоти n_x	3	10	13	11	30	39	n=106
Добуток чисел пропорційності $\Psi=k_{Ai}\cdot k_{Bi}$	1·1=1	1·3=3	1·4=4	3·1=3	3·3=9	3·4=12	
$a=\frac{n_x}{\psi}$	3,0	3,3	3,3	3,7	3,3	3,3	a_{min}=3
$\psi \cdot a_{min} = n_x^*$	1·3=3	3·3=9	4·3=12	3·3=9	9·3=27	12·3=36	n^*=96
	1:3:4			1:3:4			

Однофакторний комплекс.

При вивченні дії на результативну ознака одного фактора, завжди присутня тільки одна пропорція частот по градаціях цього чинника, яку ні з чим порівнювати. Тому для однофакторних комплексів відпадає вимога пропорційності або рівномірності: однофакторні комплекси ортогональні при будь-якому співвідношенні частот по градаціях фактора. До однофакторном комплексам повною мірою відноситься вимога рандомізації.

Двохфакторні і багатофакторні комплекси.

У двохфакторних (багатофакторних) комплексах необхідно обов'язково мати незалежність досліджуваних факторів і, бажано, пропорційність у частотах. Як і всі інші комплекси, двохфакторні повинні бути рандомізовані.

Вивчаючи дію більше одного фактора, необхідно враховувати не тільки вплив кожного фактора окремо, але і їх поєднань, як було показано в прикладі вище, тобто повинні бути розраховані приватні середні по фактору А, В, по сполученням факторів АВ і за їх загальною сумарною дією - за фактором *x*. По кожному ряду середніх розраховані центральні відхилення, сума квадратів яких дає дисперсію по кожному фактору.

7 РЕГРЕСІЙНИЙ АНАЛІЗ

Спочатку розглянемо кілька важливих понять регресійного аналізу.

Регресією називається зміна функції (Y) при певних змінах одного або декількох аргументів (x).

Функція - це ознака, що залежить від інших ознак - аргументів. Залежність функції від аргументів може бути:

- фізіологічною;
- умовно прийнятої в дослідженні.

Прикладом фізіологічної залежності може служити залежність ваги тварини (функції) від його віку (аргументу).

Якщо по довжині визначається вага тварини, вважається що вага залежить від довжини, якщо необхідно передбачити розміри тварин різної ваги, то приймається, що довжина залежить від ваги. Це приклад умовної залежності. Розкрити функцію - означає знайти закономірності, за якими змінюється досліджувана ознака залежно від зміни однієї або кількох ознак.

Якщо зміни функції досліджується в залежності від однієї ознаки, то регресія називається простою:

$$Y = f(x) \Leftrightarrow x = f(y)$$

Якщо вивчається залежність зміни функції від зміни декількох ознак, регресія називається множинною.

$$Y = f(x_1, x_2, ..., x_n)$$

Якщо при однаковому прирощені аргументу, але при різних його значеннях (малих, великих або середніх) функція має неоднакове прирощення, причому середня її зміна не йде по прямій, то регресія називається криволінійною (рис. 7.1).

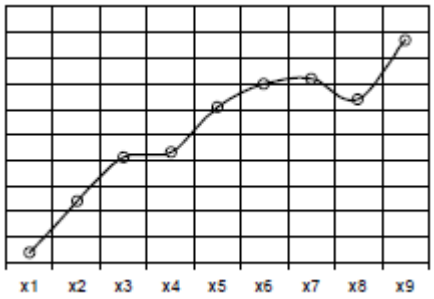


Рисунок 7.1 - Криволінійна регресія

$$\Delta x_1 = \Delta x_2 = \ldots = \Delta x_n;\ \ \Delta y_1 \neq \Delta y_2 \neq \ldots \neq \Delta y_n;\ \ \frac{\Delta y_1}{\Delta x_1} \neq \frac{\Delta y_2}{\Delta x_2} \neq \ldots \neq \frac{\Delta y_n}{\Delta x_n} \neq \text{Const}$$

Якщо при будь-якому значенні (малому, середньому або великому) аргументу однакова зміна його приводить до однакової зміни значення функції, то регресія називається прямолінійною (рис.7.2).

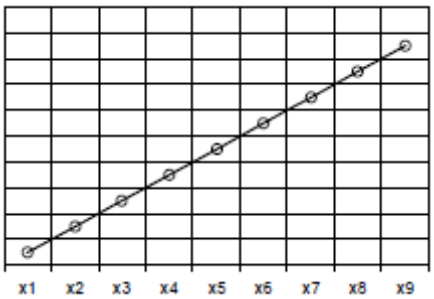


Рисунок 7.2 - Прямолінійна регресія

$$\Delta x_1 = \Delta x_2 = \ldots = \Delta x_n;\ \ \Delta y_1 = \Delta y_2 = \ldots = \Delta y_n;\ \ \frac{\Delta y_1}{\Delta x_1} = \frac{\Delta y_2}{\Delta x_2} = \ldots = \frac{\Delta y_n}{\Delta x_n} = \text{Const}$$

На практиці регресію прийнято зображати у вигляді: регресійного ряду (емпіричного чи теоретичного); лінії регресії (емпіричної чи теоретичної); коефіцієнтів регресії, які утворюють рівняння регресії, наприклад:

$$Y = \sum_{i=0}^{n} a_i \cdot x^i = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_n \cdot x^n$$

Розглянемо кожен із способів подання регресії.

Емпіричний ряд регресії - це подвійний ряд цифр, який включає значення аргументу і відповідні їм значення функції, які отримані дослідним шляхом. Приклад емпіричного ряду регресії надано в табл.7.1.

Таблиця 7.1 - Приклад емпіричного ряду регресії

Вік (x), роки	2	3	4	5	6	7	8	9	10
Жива вага (Y), кг	394	414	420	433	451	460	462	454	477

Складання емпіричного ряду регресії. Для складання емпіричного ряду регресії весь первинний матеріал розбивається на стільки груп, скільки встановлено градацій аргументу, і по кожній групі підраховується ΣV - загальна сума значень функції і n - число особин. Середня виходить простим діленням першого числа на друге:

$$M_i = \frac{\sum V_i}{n}$$

При графічному зображенні емпіричного ряду регресії - аргумент, наприклад вік, відкладається по осі абсцис, а функція, наприклад вага, відкладається по осі ординат. У результаті отримують емпіричну лінію регресії (див. рис. 7.3).

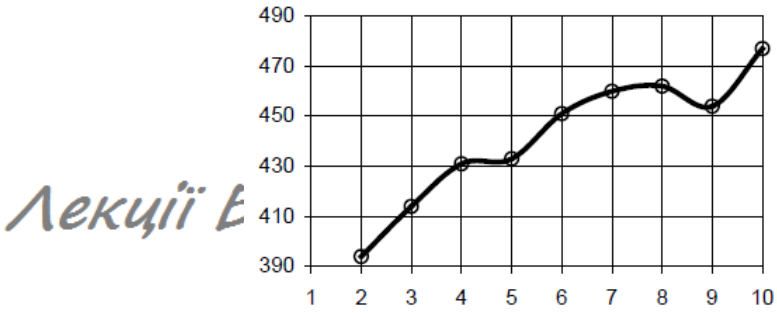


Рисунок 7.3 - Емпірична лінія регресії

Траєкторія емпіричної лінії регресії майже ніколи не буває плавною: у межах одних інтервалів аргументу, функція має підвищене, інших - знижене, а іноді і негативне прирощення, що на графіку дає ламану криву.

Ламаний характер емпіричної лінії регресії відображає звичайну невиравненість загальних умов розвитку ознаки - функції (ваги) на різних ділянках зміни ознаки - аргументу (віку). Якщо вивчається регресія ваги за віком, то стає очевидним, що вплив всіляких агентів на вікові зміни ваги не залишається однаковим впродовж періоду зростання. В одному віці вся сума впливів складається в комплекс, більш сприятливий для зростання, в іншому віці цей комплекс впливів не сприяє достатньому приросту ваги.

Таким чином, по виду емпіричної лінії регресії завжди можна встановити на яких ділянках зміни аргументу ознака - функція розвивався в кращих, а на яких в гірших умовах. Аналіз емпіричної лінії регресії завжди дає практично цінну характеристику всіх обставин, які пов'язані із залежністю досліджуваної функції від обраного аргументу.

На практиці, для знаходження основних форм залежності функції від аргументу необхідно з'ясувати таке течіння функції при рівномірному зміні аргументу, яке відповідає усередненому, тобто однаковому впливу всього комплексу умов, які визначають розвиток ознаки-функції.

Знаходження усередненого вирівняного течіння функції, в деякій мірі, подібно визначенню середньої арифметичної декількох значень ознаки.

Середня арифметична виходить шляхом згладжування індивідуальних відмінностей усереднення ознаки, крім цього вона близько стоїть до всіх індивідуальним значень, так що сума квадратів відхилень від їх середньої є величина найменша. Ці ж принципи покладені і в основу знаходження усередненого течії функції.

Однак між усереднюванням течії функції і визначенням середньої арифметичної є й істотні відмінності: середня арифметична завжди має справу з однією змінною величиною (від особини до особини), а вирівняні течією функції завжди має справу з двома або декількома змінними величинами, з яких одна (функція) величина залежить від інших (аргументи) величин.

Процес отримання усередненої течії зміни функції при рівномірному збільшенні значення аргументу - називається вирівнюванням емпіричних рядів.

У результаті вирівнювання на основі емпіричної ламаної лінії виходить усереднена, плавна теоретична лінія регресії, яка відображає основну закономірність залежності функції від аргументу.

На практиці вирівнювання емпіричних рядів виробляється графічно або аналітично.

При аналітичному методі вирівнювання емпіричних рядів в результаті складання рівняння регресії спочатку розкривається форма залежності даної функції від обраного аргументу. Підставляючи в отримане рівняння регресії послідовні значення аргументу, можна визначити теоретичний ряд значень функції, а наносячи ці значення на графік - отримати теоретичну лінію регресії .

Для емпіричного ряду, який був розглянутий у прикладі вище, рівняння регресії має наступний вигляд:

$$y = a - b \cdot 10^{-c \cdot x} = 470 - 72 \cdot 10^{-0,1312 \cdot x}$$

де: y - теоретична вага особини; x - вік у роках;

a = 470 - максимальне значення ваги, до якого асимптотично наближається дана функція в міру збільшення аргументу (віку);

b = 72 - сума приросту від першого наявного значення віку до його значення при зупинці росту;

c = 0,1312 - показник темпу зростання.

Теоретичний і емпіричний ряди наведені в наступній табл.7.2.

Таблиця 7.2

Вік (y), роки	2	3	4	5	6	7	8	9	10
Емпірічний ряд									
Жива вага (x), кг	394	414	420	433	451	460	462	454	477
Теоретичний ряд									
Жива вага (x). кг	398	417	431	441	449	454	458	461	464

Як відомо з кореляційного аналізу, в простому випадку, при прямолінійній регресії:

$$y=a\cdot x$$

залежність функції від аргументу може бути виражена одним числом - коефіцієнтом регресії, який показує в якому напрямку і наскільки змінюється функція при збільшенні значення аргументу на одну одиницю виміру. У природі існує безліч явищ, які обумовлені безліччю причин. Тому існує дуже багато форм залежності функцій від різних аргументів.

Дослідження цих форм, виражених математичними рівняннями, становить основний зміст вчення про регресії ознак.

Вирівнювання емпіричних рядів регресії має велике і різнобічне застосування. Розкриваючи усереднений перебіг функції, дослідник виявляє ту закономірність досліджуваного явища, яка в емпіричному ряду була розкрита випадковостями свого прояву. Ця закономірність, виражена формулою або теоретичним рядом регресії, допомагає більш точно, з меншими помилками дати опис зовнішніх проявів закономірності, що, у свою чергу, може допомогти знаходженню і внутрішніх факторів, керуючих даним явищем. У цьому і заключається пізнавальне значення досліджень регресії різних ознак у еколого-біологічних об'єктів.

Результати цих досліджень мають також широке застосування і в практиці. Кожен вирівняний ряд дає можливість визначити значення функції при будь-якому значенні аргументу (або декількох аргументів). Ця обставина дає можливість використовувати ряди і рівняння регресії при визначенні значень таких ознак, безпосереднє вимірювання яких в звичайних умовах або неможливо, або важко.

У практичних роботах використання рівнянь і ліній регресії набуло широкого поширення при визначенні без зважування, шляхом вимірювання, нормального живої ваги тварин та їх забійної ваги за життя, ваги сіна в стогах, ваги овочів в овочесховищах, ваги силосної маси в силосах, ваги деревини в стовбурах і штабелях та ін.

Широке практичне застосування в багатьох галузях виробництва знаходить також спеціальна форма ліній регресії - номограма.

7.1 Загальні способи вирівнювання емпіричних рядів

До загальних способів вирівнювання емпіричних рядів відносяться:
- графічний спосіб;
- спосіб ковзної середньої (простий і зваженої);
- метод (спосіб) найменших квадратів (МНК .

7.1.1 Графічний спосіб

Графічний спосіб дає можливість з достатнім наближенням отримати теоретичну лінію, а потім і теоретичний ряд регресії без будь-яких обчислень.

Найбільш простим виявляється застосування графічного способу до прямолінійної регресії. У цих випадках на графік наноситься спочатку емпірична лінія регресії, потім між крайніми виступами ламаної емпіричної лінії проводиться пряма таким чином, щоб сума відстаней теоретичної прямої від точок емпіричної лінії була б найменшою.

При відомій навичці це можна зробити від руки. При цьому може допомогти і натягнута нитка або прозора лінійка з нанесеною прямий рисою. Натягнута нитка розташовується по середній течії емпіричної лінії, і після знаходження найкращого положення нитки на графіку відзначаються дві крайні точки: для мінімального і максимального значення аргументу. Теоретичною лінією регресії буде пряма, що з'єднує ці дві точки.

За теоретичної прямої можна визначити числові значення функції (ординати), що відповідають певним значенням аргументу (абсциси).

Якщо регресія не може вважатися прямолінійною, то графічне вирівнювання емпіричної кривої також може бути проведено, но для цього необхідно мати уявлення про загальні закономірності зміни функції. Наприклад, при вивченні вікових змін живої ваги сільськогосподарських тварин потрібна враховувати, що жива вага, збільшуючись з віком, поступово наближається до деякого максимального значення, після чого приріст припиняється і значення його залишається приблизно на одному максимальному рівні.

7.1.2 Спосіб ковзної середньої

Якщо форма функції невідома, то згладити злами емпіричної кривої можна, застосувавши спосіб простий ковзної середньої. Цей спосіб полягає в тому, що для кожного значення аргументу береться середня арифметична з декількох (сусідніх) значень функції .

Якщо змінна середня береться за трьома значенням аргументу, то складаються значення функції для меншого значення аргументу, для даного і для більшого. Приватне від розподілу цієї суми на 3 дає вирівняні значення функції для даної величини аргументу.

Вирівнювання емпіричного ряду методом простої ковзної середньої показано в табл.7.3. Вирівняна цим методом крива дана на рис. 7.4.

Вирівнювання емпіричних рядів способом простий ковзної середньої застосовується, коли не потрібно особливої точності і коли є достатньо довгий ряд і можна знехтувати втратою двох значень функції, відповідних крайнім значенням аргументу.

Таблиця 7.3 - Вирівнювання емпіричного ряду за способом простий і зваженою ковзної середньої

Аргумент - зміст перетравного білка (%) в раціоні телят до шестимісячного віку, функція (у) - вага телят (кг) у віці шести місяців

Відсоток білка в раціоні	Жива вага (у)	Сума трьох сусідніх (у)	Вирівняні значення y_v за методом	
			«Простий»	«Зважений»
-	y_{+2}=0,5·(2·89,5+103-125)=78,5 - додаткове значення			
-	y_{+1}=0,5·(2·103+120-147)=89,5 - додаткове значення			
56	y_1=103	-	-	130,45
53	y_2=120	348*	116	117,25
50	y_3=125	392*	131	127,6
47	y_4=147	411	137	138,9
44	y_5=139	439	146	142,8
41	y_6=153	439	146	148,5
38	y_7=147	454	151	149,5
35	y_8=154	455	152	152,0
32	y_9=154	457	152	152,8
29	y_{10}=149	462	154	151,6
26	y_{11}=159	448	149	152,0
23	y_{12}=140	451	150	144,9
20	y_{13}=152	410	137	139,75
17	y_{14}=118	-		124,85
-	y_{n+1}=0,5·(2·118+152-159)=114,5- додаткове значення			
-	y_{++2}=0,5·(2·114,5+118-140)=103,5 - додаткове значення			

*$y_1+y_2+y_3$=103+120+125=348

$y_2+y_3+y_4$=120+125+147=392

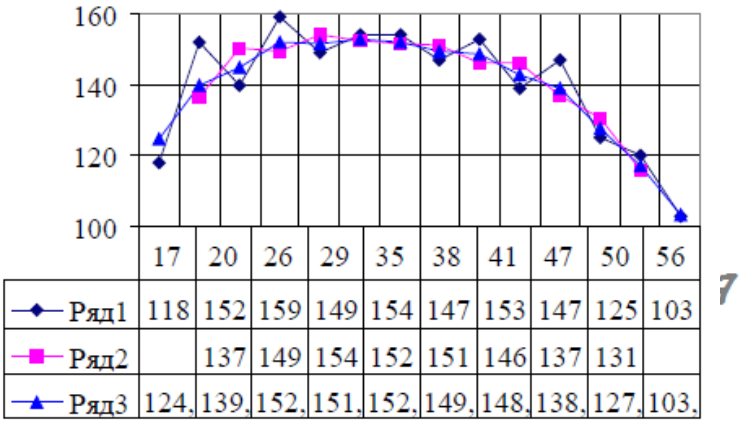


Рисунок 7.4 - Вирівнювання емпіричного ряду за способом простий і зваженою ковзної середньої

Більш точні і не пов'язані з втратою крайніх значень результати виходять при використанні зваженої ковзної середньої. При цьому способі з обох кінців ряду додається по два значення - по два члена ряду. Визначаються вони таким чином.

Перше (від кінця) значення ряду (y_1) множиться на 2, до отриманого добутку додається друге значення (y_2) , третє (y_3) пропускається, а із суми віднімається четверте (y_4) . Отримане число ділиться на 2 . Приватне буде першим додатковим значенням: y_{+1} (для початку ряду) або y_{n+1} (для кінця ряду). Всі ці дії можна виразити наступною формулою:

$$y_{+1} = \frac{2 \cdot y_1 + y_2 - y_4}{2}; \qquad y_{n+1} = \frac{2 \cdot y_n + y_{n-1} - y_{n-3}}{2}$$

Другі додаткові значення з обох кінців ряду:

$$y_{n+2} = \frac{2 \cdot y_{+1} + y_1 - y_3}{2}; \qquad y_{n+2} = \frac{2 \cdot y_{n+1} + y_n - y_{n-2}}{2}$$

Для його розрахунку потрібно використовувати перше додаткове значення, а також перше і третє значення початкового емпіричного ряду. Для розглянутого прикладу в двох верхніх і в двох нижніх рядках табл.7.3 показано отримання додаткових значень.

Для коротких рядів додаткові значення можна отримувати, користуючись такими формулами:

$$y_{+1} = \frac{4 \cdot y_1 + y_2 - 2 \cdot y_3}{3}; \qquad y_{+2} = \frac{4 \cdot y_{+1} + y_1 - 2 \cdot y_2}{3}$$

Після встановлення додаткових значень приступають до вирівнювання емпіричного ряду. Вирівняні значення виходять шляхом обчислення зваженої середньої арифметичної з п'яти сусідніх емпіричних значень функції, взятих відповідно з вагами 1; 2; 4; 2; 1.

Для того, щоб отримати, наприклад, перше вирівняне значення функції, потрібно суму другого додаткового, подвоєного першого додаткового, почетвереного першого емпіричного, подвоєного другого емпіричного і третього емпіричного значень функції розділити на суму ваг (1+2+4+2+1 = 10).

Це можна виразити наступною формулою:

$$y_{v1} = \frac{y_{+2} + 2 \cdot y_{+1} + 4 \cdot y_1 + 2 \cdot y_2 + y_3}{10}$$

Для розглянутого емпіричного ряду першого вирівняні значення функції будуть рівні:

$$y_{v1} = \frac{78{,}5 + 2 \cdot 89{,}5 + 4 \cdot 103 + 2 \cdot 120 + 125}{10} = 103{,}45$$

$$y_{v2} = \frac{89{,}5 + 2 \cdot 103 + 4 \cdot 120 + 2 \cdot 125 + 147}{10} = 117{,}25$$

У табл.7.3 наведено розрахунок всіх вирівнюються значень функції для розглянутого прикладу. Ряд живої ваги шестимісячних телят, вирівняний методами простий і зваженою ковзної середньої, показаний на рис. 7.4.

7.1.3 Метод найменших квадратів (МНК)

Найбільш поширеним загальним аналітичним способом вирівнювання емпіричних рядів регресії є метод (спосіб) найменших квадратів (МНК). Цей метод надає найбільш універсальну можливість для вирівнювання та визначення виду аналітичної функції (прямолінійною, зворотного, параболічної, гіперболічної, статечної, логарифмічною, експоненційної, періодичної, простий, множинною та комбінації їх) наближено замінює табличні дані отримані експериментальним шляхом.

МНК призначений для вибору із сукупності призначеного типу кривих (прямолінійною, зворотною, параболічної і т.д.) такої кривої, для якої сума квадратів відхилень емпіричних даних від вирівняних (обчислених за формулою даного типу кривої) є найменшою, тобто:

$$(y_e - y_p)^2 \Rightarrow \min$$

Розглянемо на прикладі визначення лінійної залежності МНК:

$$y = b \cdot x + a$$

З умови мінімуму, для цієї залежності необхідно підібрати b і a так, щоб відхилення експериментальних точок від теоретичної кривої було б найменшим, тобто:

$$y_{Pi} = b \cdot x_i + a$$

$$\delta_i = y_{Эi} - y_{Pi}$$

$$y_{Pi} = y_{Эi} - \delta_i = b \cdot x_i + a$$

$$\delta_i = y_{Эi} - (b \cdot x_i + a)$$

Будемо шукати функцію з невідомими b і a, у якої сума δ_i була б найменшою позитивною величиною. При цьому потрібно враховувати, що можливі й негативні значення δ_i, тому будемо шукати екстремум мінімуму для параболи, тобто зведемо в квадрат обидві частини рівняння різниці:

$$(\delta_i)^2 = (y_{эi} - (b \cdot x_i + a))^2$$

Сума квадратів відхилень експериментальних даних від обчислених складе:

$$S = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (y_{эi} - (b \cdot x_i + a))^2$$

Складемо систему рівнянь пошуку екстремуму (мінімуму):

$$\begin{cases} \frac{\partial S}{\partial b} = 0 \\ \frac{\partial S}{\partial a} = 0 \end{cases}, \qquad \begin{cases} \sum_{i=1}^{n} 2 \cdot (y_{эi} - (b \cdot x_i + a)) \cdot (-x_i) = 0 \\ \sum_{i=1}^{n} 2 \cdot (y_{эi} - (b \cdot x_i + a)) \cdot (-1) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} x_i \cdot y_{эi} - b \cdot \sum_{i=1}^{n} x_i^2 - a \cdot \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_{эi} - b \cdot \sum_{i=1}^{n} x_i - a \cdot n = 0 \end{cases}$$

Щоб вирішити цю систему рівнянь спочатку щодо b, помножимо перше рівняння на n, а друге – на $\left(-\sum_{i=1}^{n} x_i\right)$:

$$n \cdot \sum_{i=1}^{n} x_i \cdot y_{эi} - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_{эi} - b \cdot \left(n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) = 0$$

Тоді:

$$b = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_{эi} - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_{эi}}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \qquad a = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_{эi} - b \cdot \sum_{i=1}^{n} x_i \right)$$

У загальному випадку вирівнювання емпіричних рядів регресії МНК здійснюється за такими етапами:

1. Визначення загального вигляду рівняння регресії. Виконується на основі попереднього біологічного аналізу процесів, що визначають протягом функції, або на основі розгляду емпіричної кривої.

2. Складання системи нормальних рівнянь. Виконується за такими правилами, показаним на наступному прикладі. Нехай вихідне рівняння $y = a + b \cdot x + c \cdot x^2$ зображується так, щоб функція (y) була в правій частині:

$$a + b \cdot x + c \cdot x^2 = y$$

Всі члени вихідного рівняння по черзі множаться на величини, які стоять поруч з шуканими коефіцієнтами a, b, c, тобто на 1, на x, на x^2:

$$(\times 1): \quad a + b \cdot x + c \cdot x^2 = y$$

$$(\times x): \quad a \cdot x + b \cdot x^2 + c \cdot x^3 = y \cdot x$$

$$(\times x^2): \quad a \cdot x^2 + b \cdot x^3 + c \cdot x^4 = y \cdot x^2$$

У кожного доданка рівняння ставиться знак підсумовування. Логічним є винести шукані коефіцієнти a, b, c за цей знак, а також слід враховувати, що $\sum_{i=1}^{n} 1 = n$ - число пар аргумент - функція:

$$a \cdot n + b \cdot \sum x + c \cdot \sum x^2 = \sum y$$

$$a \cdot \sum x + b \cdot \sum x^2 + c \cdot \sum x^3 = \sum y \cdot x$$

$$a \cdot \sum x^2 + b \cdot \sum x^3 + c \cdot \sum x^4 = \sum y \cdot x^2$$

Отримані рівняння і є система нормальних рівнянь для даної вихідної параболічної функції . Ці правила застосовні і для будь-якої вихідної формули, наприклад:

$$a + \frac{b}{x} = y$$

$$(\times 1): \quad a + b \cdot x + c \cdot x^2 = y \qquad a \cdot n + b \cdot \sum \frac{1}{x} = \sum y$$

$$\left(\times \frac{1}{x}\right): \quad \frac{a}{x} + \frac{b}{x^2} = \frac{y}{x} \qquad a \cdot \sum \frac{1}{x} + b \cdot \sum \frac{1}{x^2} = \sum \frac{y}{x}$$

3. Визначення числового значення сум, що входять в нормальні рівняння. Проводиться шляхом підсумовування попередньо обчислених рядів:

$$\sum x; \quad \sum x^2; \quad \sum x^3; \quad \sum x^4; \quad \sum y; \quad \sum y \cdot x; \quad \sum y \cdot x^2$$

4. Визначення коефіцієнтів основного рівняння. Проводиться шляхом рішення системи нормальних рівнянь звичайними алгебраїчними прийомами. У розглянутому прикладі (вихідне рівність $y = a + b \cdot x + c \cdot x^2$) коефіцієнтами будуть a, b, c.

Вирівнювання емпіричних рядів регресії МНК можна показати на наступних прикладах.

7.1.3.1 Прямолінійні функції виду $y = b \cdot x + a$

Приклад. У деяких випадках можна виконати визначення віку корів за кількістю кілець на рогах. Зв'язок між числами кілець на рогах і віком виникає тому, що кожне отелення, що відбувається зазвичай щорічно, залишає на рогах корови кільце, що відбиває уповільнення зростання роги в періоди глибокої тільності, коли головна маса живильних речовин витрачається на харчування плоду. Тому якщо до середнього числа років, що минули до першого отелення додати число кілець, помножене на середній межотельний період, то це і буде зразковим віком корови. Це можна виразити рівнянням прямої:

$$y = b \cdot x + a,$$

де а - число років до першого отелення;
х - число кілець на рогах;
у - вік корови в роках;
b - середній межотельний період.

Знаючи вихідне рівняння, можна скласти систему нормальних рівнянь. В даному випадку вона буде досить простою:

$$a \cdot n + b \cdot \sum x = \sum y$$

$$a \cdot \sum x + b \cdot \sum x^2 = \sum y \cdot x$$

або скористатися раніше отриманими формулами для розрахунку а і b:

$$b = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$a = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_i - b \cdot \sum_{i=1}^{n} x_i\right)$$

Таким чином, необхідно визначити наступні чотири величини сум:

$$\sum x; \quad \sum x^2; \quad \sum y; \quad \sum y \cdot x$$

Ці суми наведені в табл.7.4.

Таблиця 7.4 - Вирівнювання емпіричного ряду регресії віку корів (у) по числу кілець на рогах (х) МНК

Знаходження параметрів y = b·x + a				Побудова теоретичного ряду		
x	y^*	x^2	yx	bx	y=b·x+a	$y-y^*$
11	13,3	121	146,3	10,955	13,4	+0,1
10	12,4	100	124,0	9,960	12,4	0,0
9	11,5	81	103,5	8,964	11,4	-0,1
8	10,5	64	84,0	7,968	10,4	-0,1
7	9,5	49	66,5	6,972	9,4	-0,1
6	8,3	36	49,8	5,975	8,4	+0,1
5	7,4	25	37,0	4,980	7,4	-0,1
4	6,5	16	26,0	3,984	6,4	-0,1
3	5,5	9	16,5	2,988	5,4	-0,1
2	4,4	4	8,8	1,992	4,4	0,0
1	3,4	1	3,4	0,996	3,4	0,0
Σ=66	Σ=92,7	Σ=506	Σ=665,8	-	-	-

$$b = \frac{11 \cdot 665{,}8 - 66 \cdot 92{,}7}{11 \cdot 506 - (66)^2} = +0{,}996$$

$$a = \frac{1}{11} \cdot (92{,}7 - 0{,}996 \cdot 66) = 2{,}45$$

Таким чином, теоретичне значення віку по числу кілець на рогах можна визначити за формулою:

$$y = 0{,}996 \cdot x + 2{,}45$$

Теоретичний ряд віку за кількістю кілець на рогах показаний на рис. 7.5.

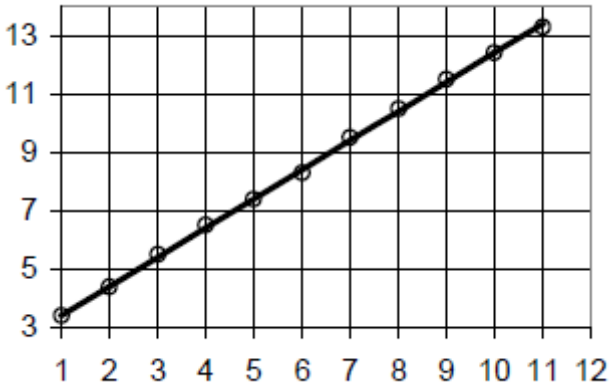


Рисунок 7.5 - Теоретичний ряд віку за кількістю кілець на рогах

Коефіцієнт кореляції для отриманого рівняння розраховуємо за формулою:

$$r = \frac{\sum_{i=1}^{n}\left[(x_i - M_x)\cdot(y_i - M_y)\right]}{\sqrt{\sum_{i=1}^{n}(x_i - M_x)^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i - M_y)^2}},$$

де M_x і M_y - середньоарифметичні відповідно для x і y:

$$M_x = \frac{\sum x}{n} = \frac{66}{11} = 6 \qquad M_y = \frac{\sum y}{n} = \frac{92{,}7}{11} = 8{,}43 \qquad r = 0{,}9997$$