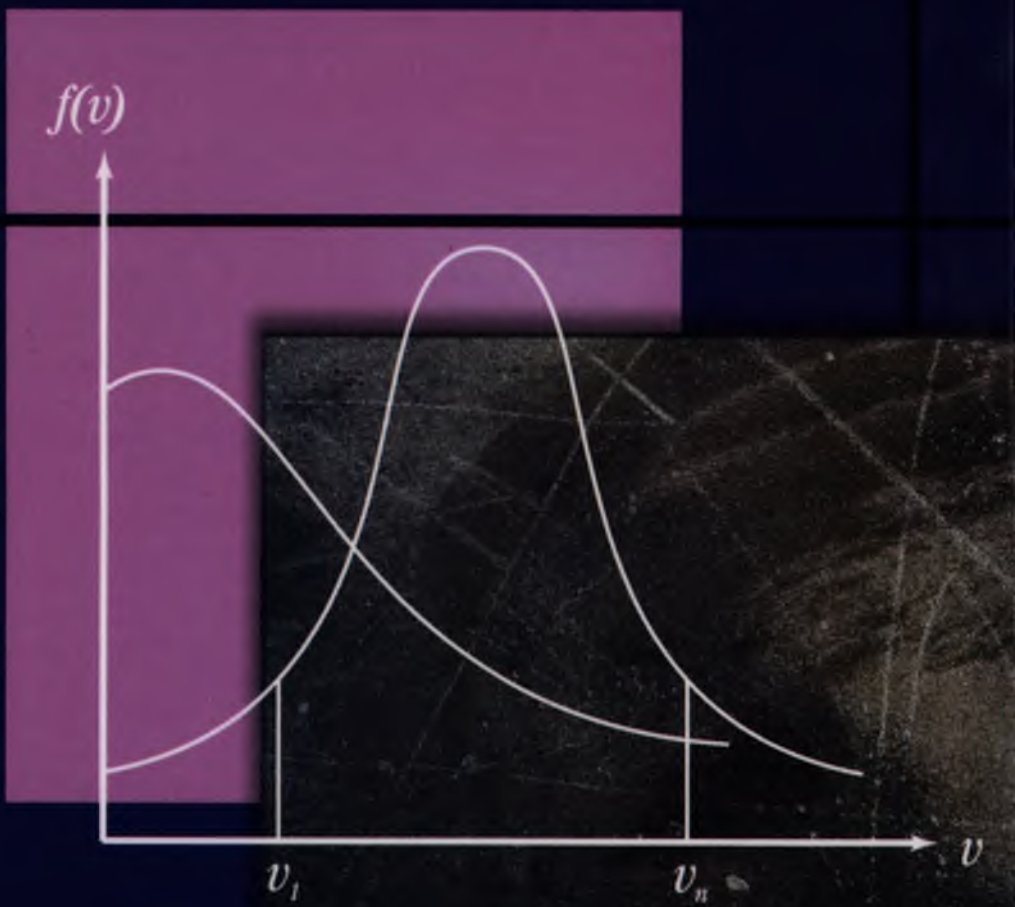
В.К. Горкавий
В.В. Ярова

МАТЕМАТИЧНА СТАТИСТИКА



МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Харківський національний аграрний університет ім. В.В. Докучаєва

В.К. ГОРКАВИЙ
В.В. ЯРОВА

МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник

Рекомендовано
Міністерством освіти і науки України
для студентів вищих навчальних закладів

Київ – 2004

УДК 631. 1 (075.8)
ББК 22.172я73
Г 67

Рецензенти:

***В.Я. Заруба**, доктор економічних наук, професор*
(Національний технічний університет «Харківський політехнічний інститут»)
***Ю.С. Коваленко**, доктор економічних наук, головний науковий співробітник*
(Національний науковий центр, Інститут аграрної економіки)
***А.Т. Опря**, доктор економічних наук, професор*
(Полтавська аграрна академія)

Рекомендовано
Міністерством освіти і науки України
як навчальний посібник для студентів вищих закладів освіти
(лист №14/18.2 – 689 від 7 квітня 2004 року)

Горкавий В. К., Ярова В. В.

Г 67 Математична статистика: Навчальний посібник. — К.: ВД «Професіонал», 2004, — 384 с.

ISBN 966-8556-44-5

Висвітлено предмет, основні категорії, метод і теоретичні основи математичної статистики, ряди розподілу та їх характеристики. Розглянуто методи обробки й аналізу статистичної інформації, зокрема вибірковий метод, перевірку статистичних гіпотез, дисперсійний і кореляційний аналіз, а також багатомірні методи обробки статистичної інформації. Значна увага приділена використанню комп'ютерних технологій у математичній статистиці. До кожного розділу розроблено перелік питань для самоконтролю знань студентів.
Для студентів економічних спеціальностей, аспірантів, викладачів і науковців вищих аграрних навчальних закладів.

УДК 631. 1 (075.8)
ББК 22.172я73

ISBN 966-8556-44-5

ЗМІСТ

Присвячується 75 річниці економічного факультету ХНАУ

ПЕРЕДМОВА

Удосконалення форм господарювання, реформування відносин власності, і переорієнтація економіки на ринок потребують відповідного кадрового забезпечення і суттєвих змін у підготовці фахівців економічного профілю.

Навчання з математичної статистики є важливою частиною методологічної підготовки економічних кадрів у питаннях якісно-кількісного аналізу масових явищ. Оволодіння сучасними методами збирання, обробки й аналізу статистичної інформації — невід'ємний елемент підготовки висококваліфікованих економістів і менеджерів.

Використання математико-статистичних методів надає змогу оцінити підприємницькі й фінансові ризики, створити умови для підвищення ефективності виробництва на основі вірогідної оцінки стану й можливостей різних сфер діяльності, своєчасного визначення тенденцій, прогнозування їх розвитку та оцінки функціонування ринкових відносин.

Відповідно до діючих державних положень вищої освіти курс математичної статистики віднесено до дисциплін нормативної частини освітньо-професійних програм підготовки бакалаврів, спеціалістів і магістрів різних економічних спеціальностей і напрямів.

Метою математичної статистики як навчальної дисципліни є формування у майбутніх менеджерів базових знань з основ застосування математико-статистичного апарата для розв'язування теоретичних і практичних економічних задач.

Завдання вивчення курсу — надання студентам знань щодо основних визначень, теорем, правил та формування умінь: виконувати якісний і кількісний математичний аналіз випадкових подій, випадкових величин та систем таких величин; проводити математичну

обробку статистичних даних; давати статистичну оцінку параметрів генеральної сукупності; здійснювати статистичну перевірку гіпотез; використовувати елементи дисперсійного аналізу, теорії кореляції; включати результати досліджень у математичні моделі економічних задач.

Студенти повинні знати: наукові основи математичної статистики і принципи організації статистичної роботи; методи збирання, обробки, зберігання і передачі статистичної інформації; традиційні та сучасні економіко-статистичні методи і прийоми аналізу економічної інформації; систему статистичних показників комплексного аналізу підприємницької діяльності суб'єктів господарювання; методику аналізу і прогнозування розвитку соціально-економічних явищ.

Студенти повинні вміти: збирати, систематизувати й аналізувати економічну інформацію з використанням сучасних електронно-обчислювальних машин, робити узагальнення, висновки та приймати кваліфіковані управлінські рішення; давати кількісну та якісну оцінку досліджуваних явищ та процесів, виявляти закономірності та тенденції їх розвитку; використовувати економіко-статистичні методи і моделі для вивчення стану і прогнозування розвитку економіки.

При підготовці навчального посібника враховано основні положення Концепції побудови національної статистики України та Державної програми переходу України на міжнародну систему обліку і статистики. Посібник складено відповідно до сучасної програми курсу з урахуванням специфіки роботи економістів і менеджерів.

Розділ 1
ПРЕДМЕТ І ТЕОРЕТИЧНІ ОСНОВИ МАТЕМАТИЧНОЇ СТАТИСТИКИ

1.1 Предмет математичної статистики

Термін «статистика» походить від латинського «status», що означає становище, стан явищ. Уперше він був уведений в ужиток у середені XVIII століття німецьким ученим, професором філософії Готфрідом Ахенвалем (1719 – 1772), який з 1746 р. у Морбургському університеті почав викладати нову навчальну дисципліну, яку і назвав статистикою.

Як окрема галузь науки статистика виникла з повсякденних практичних потреб людей, оскільки для державного управління потрібна інформація про чисельність населення, наявність і склад земель, стан торгівлі тощо. Об'єктом її вивчення є соціальні, економічні, політичні та культурні явища і процеси суспільного життя.

Статистика вивчає кількісну сторону масових суспільних явищ і процесів у нерозривному зв'язку з їх якісною стороною, досліджує кількісний вираз закономірностей суспільного розвитку у конкретних умовах місця і часу. Вона вивчає також вплив природних, технічних, соціальних та економічних факторів на умови й результати виробництва, зворотний вплив розвитку суспільного виробництва на умови життя людей. Так, статистика вивчає закономірності розвитку економіки, кількість і склад населення країни, кількість і рівень використання трудових, матеріальних і фінансових ресурсів, доходи та матеріальне становище населення, рівень і динаміку цін на засоби виробництва та предмети споживання, розвиток народної освіти, культури й охорони здоров'я, природні ресурси й охорону навколишнього середовища, стан і розвиток матеріального виробництва тощо.

Статистика — багатогалузева самостійна наука, яка пройшла складний шлях становлення. У процесі розвитку в її складі виділилися математична статистика, загальна теорія статистики, економічна та соціальна статистики, галузеві статистики (промислова,

сільськогосподарська, транспортна, торговельна, медична, правова і т.д.).

Математична статистика — це галузь наукових знань, присвячених математичним методам систематизації, обробки й аналізу даних спостереження масових явищ з метою виявлення та використання в науково-практичній діяльності характерних для них статистичних закономірностей.

Застосування математичних методів у статистиці розширює та поглиблює кількісно-якісний аналіз природних і суспільних масових явищ. На відміну від математики в статистиці кількісні характеристики не абстрактні, а пов'язані з якісним змістом масових явищ, що стосуються конкретних умов місця та часу.

Одержані за допомогою методів математичної статистики кількісні характеристики досліджуваних масових явищ дозволяють мати об'єктивне та всебічне уявлення про характер причинно-наслідкових зв'язків між ними й одержувати надійні параметри для обгрунтування напрямів соціально-економічної політики, оцінки результатів і прогнозування розвитку економіки.

Математична статистика є самостійною науковою дисципліною, що має свій предмет і метод дослідження. Предметом математичної статистики є формальна математична сторона статистичних методів дослідження, не пов'язана зі специфічною природою об'єктів, які вивчаються.

Як самостійна наука математична статистика (в ряді літературних джерел її називають прикладною статистикою) розробила та використовує потужний арсенал прийомів і способів обробки масових даних, одержаних у результаті наукових і практичних досліджень, які носять імовірносно-випадковий характер. Невипадково сучасну математичну статистику визначають як науку про прийняття рішень в умовах невизначеності.

Область застосування математико-статистичних методів обмежується вимогами, суть яких зводиться до того, щоб досліджувані

явища були підпорядковані достатньо визначеним статистичним закономірностям. Використання методів математичної статистики дозволяє за результатами обмеженої кількості спостережень масових явищ скласти уявлення про ті вірогідні закономірності, яким підпорядковані ці явища і передбачити умови їх прояву в майбутньому.

Найважливішими розділами математичної статистики є статистична оцінка параметрів розподілу, статистична перевірка гіпотез, дисперсійний, регресійний, кореляційний, коваріаційний, факторний, кластерний та інші багатомірні методи аналізу, а також статистичні методи аналізу випадкових процесів. Важливою частиною сучасної математичної статистики є методи планування експериментів, поступового аналізу досліджень, контролю якості і надійності.

Основними завданнями математичної статистики є: розробка способів збирання, систематизації і використання статистичної інформації, одержаної у результаті спостереження або спеціально виконаних експериментів для формулювання наукових і практичних висновків; визначення узагальнюючих характеристик і міри варіації досліджуваних явищ, установлення законів розподілу випадкових змінних, одержаних у результаті спостереження, оцінка невідомої імовірності події, невідомої функції та параметрів розподілу; статистична перевірка гіпотез відносно параметрів або форми розподілу за допомогою спеціальних критеріїв; визначення та кількісна оцінка взаємозв'язків між явищами та їх істотності.

Перші дослідження в галузі математичної статистики належать швейцарським ученим, братам Якобу (1654 – 1705) і Йоганну (1667 – 1748) Бернулі, які спільно розробили методи оцінки варіації ознак. Подальший розвиток математичної статистики у XIX сторіччі, пов'язаний з іменами П. Л. Чебишева, А.А. Маркова, О. М. Ляпунова, К. Гауса, А. Кетле, Ф. Гальтона, К. Пірсона та ін.

У XX сторіччі найбільший внесок у розвиток математичної статистики зробили В.І. Романовський, Є.Є Слуцький, С. Н. Берштейн, А. М. Колмогоров, Ю.В. Лінник, М. В. Смирнов, В. Госсет (Стьюдент), Р. Фішер, Е. Пірсон, Ю. Нейман, А. Вальд.

1.2 Основні категорії математичної статистики

Математична статистика досліджує масові явища. У зв'язку з цим вона вивчає такі категорії, як статистична сукупність, варіація, варіююча ознака, статистична закономірність, статистичний показник.

Статистична сукупність — це сума об'єктів (подій, елементів, явищ тощо), які мають єдину якісну основу, але відрізняються певними ознаками. Найважливішою особливістю статистичної сукупності є однорідність, одноякісність її елементів. Так, для сільськогосподарських підприємств характерним є те, що всі вони займаються виробництвом продукції рослинництва і тваринництва. Ця спільність і об'єднує їх у сукупність сільськогосподарських підприємств. В окремих державних, приватних, фермерських та інших господарствах статистичними сукупностями можуть бути чисельність працівників, кількість виробничих підрозділів, посівні площі, поголів'я тварин тощо. Окремі об'єкти (елементи, явища і т. ін.), які становлять статистичну сукупність, називаються одиницями сукупності. Вони характеризуються індивідуальними ознаками, на підставі яких статистика узагальнює дані сукупності в цілому. Передумовою такого узагальнення є вимірювання, тобто характеристика явищ через числові значення.

Маючи узагальнені характеристики масових явищ, статистика за їх допомогою виявляє властиві досліджуваним сукупностям закономірності.

Статистична закономірність — це повторюваність, послідовність і порядок у явищах. Вона характеризує кількісну закономірність зміни у просторі або часі масових явищ і процесів. Виявляється статистична закономірність тільки в масовому процесі при великій кількості одиниць сукупності. При цьому вплив другорядних факторів має випадковий характер і в цілому взаємно погашається.

Статистичні закономірності масових явищ відображають характер дії об'єктивних законів розвитку в конкретних умовах часу та простору. За змістом їх поділяють на:

1) закономірності розвитку (динаміки) явищ (збільшення кількості фермерських господарств, підвищення продуктивності праці);
2) закономірності зміни структури явищ (зменшення частки сільськогосподарського населення у загальній його чисельності, підвищення частки господарств населення у загальному обсязі валової продукції сільського господарства);
3) закономірності розподілу одиниць сукупності (розподіл населення за рівнем середньодушового сукупного доходу, розподіл безробітних за віком і освітою);
4) закономірності зв'язку між явищами (залежність продуктивності праці від рівня енергоозброєності, залежність продуктивності тварин від рівня їх годівлі).

Для статистичної сукупності характерна зміна значень ознак, що вивчаються. **Ознаки** — це властивості, характерні риси або інші особливості одиниць сукупності, на підставі яких можна зробити висновок про стан певного явища. Так, ознаками сільськогосподарських підприємств є земельна площа, чисельність працівників, розмір основних і оборотних фондів, обсяг виробленої продукції тощо. Ознаками працівників є стать, стаж роботи, освіта, професія, кваліфікація і т.д.

Значення ознак одиниць сукупності змінюється від елемента до елемента або від одного періоду до іншого. Ознака, яка має в межах сукупності різні значення, називається **варіюючою**, а коливання значень ознаки — **варіацією**. Склад елементів і спосіб сполучення їх характеризують структуру сукупності.

Індивідуальна роль ознаки в сукупності визначається як загальними для всіх одиниць причинами, так і конкретними особливостями кожної одиниці. У зв'язку з цим розмір ознаки в окремих одиниць сукупності варіює і масове явище характеризується численністю значень ознаки.

Предметом математичної статистики є особливі ознаки стану і розвитку масових явищ. Такі ознаки дістали назву об'єктивних статистичних показників.

Статистичний показник — це загальна істотна ознака якогонебудь масового явища у її якісній і кількісній визначеності для конкретних умов місця та часу. Кожний статистичний показник має кількісний вираз. Разом з тим кількість у статистиці завжди має відповідну якість. Як єдність кількості і якості статистичні показники характеризують міру явища.

Всебічне вивчення складних масових явищ забезпечується тільки при використанні системи статистичних показників, яка становить їх взаємопов'язану сукупність, що характеризує певне явище. Наприклад, для вивчення масових явищ і процесів, що відбуваються у сільськогосподарському виробництві, застосовують систему статистичних показників, які характеризують соціально-економічні та природні умови, рівень, темпи й ефективність його розвитку. Ця система включає насамперед показники обсягу та якості наявних ресурсів сільськогосподарського виробництва (землі, робочої сили, основних і оборотних виробничих засобів, їх співвідношення і використання).

Наслідком використання виробничих ресурсів є формування витрат і результатів сільськогосподарського виробництва. У зв'язку з цим статистичні показники характеризують також витрати за видами продукції і окремими елементами, обсяг валової, товарної і чистої продукції.

Крім того, статистичні показники характеризують ефективність використання основних елементів сільськогосподарського виробництва (землі, тварин, праці та засобів виробництва) і відображують інтенсивні умови господарської діяльності. Їх обчислюють через співвідношення продукції. витрат і ресурсів. Найважливішими показниками ефективності виробництва є вихід валової, товарної та чистої продукції з розрахунку на одиницю земельної площі, одиницю виробничих засобів, одиницю робочої сили та праці, одиницю витрат виробництва.

Окрему групу становлять показники матеріально-технічного постачання ресурсів для відшкодування поточних витрат, ремонтів, реновації елементів основних засобів і розширення виробництва.

За формою та змістом статистичні показники поділяють на абсолютні і відносні. **Абсолютні показники** характеризують розміри суспільних явищ, наприклад, розмір посівних площ, кількість тварин, обсяг продукції, суму прибутку тощо. Вони відповідають на запитання «скільки» і завжди виражаються іменованими числами (гектарами, центнерами, кілометрами, метрами, тоннами і т.ін.). **Відносні показники** виражають кількісні співвідношення між абсолютними показниками, а абсолютні показники — кількість одиниць і загальні розміри явищ у всіх одиниць сукупності.

Загальний розмір явищ (земельної площі, поголів'я тварин, чисельності працюючих, обсягу продукції тощо) для всієї сукупності в статистиці називають **обсягом явища**, а кількість одиниць, які мають певну ознаку (кількість підприємств, рослин, тварин, працівників) — обсягом сукупності. Якщо, наприклад, вивчається поголів'я корів у фермерських господарствах району, то в цьому разі загальне поголів'я корів в усіх фермерських господарствах буде обсягом явища, а кількість господарств — обсягом сукупності; якщо ж досліджується продуктивність корів у фермерських господарствах, то в цьому разі валове виробництво молока буде обсягом явища, а поголів'я корів — обсягом сукупності.

1.3 Метод і теоретичні основи математичної статистики

Під методом будь-якої науки розуміють спосіб дослідження наукою явищ природи і процесів суспільного життя. Загальним методом пізнавання для всіх наук, у тому числі для статистики, є матеріалістична діалектика. Згідно з її принципами всі явища природи і суспільного життя органічно взаємопов'язані та взаємозумовлені. Вони залежать одне від одного, впливають одне на одне і перебувають у постійному русі та розвитку. З окремих явищ формується єдність їх складових частин і властивостей.

Статистика вивчає суспільні явища не ізольовано одне від одного, а у взаємозв'язку та взаємозалежності, виділяє різні типи і форми суспільних явищ, оцінює вплив факторів, які формують динаміку та варіацію явищ, розкриває процес переходу кількісних змін у якісні.

Спираючись на загальні принципи діалектичного методу, статистика розробила свої, властиві їй прийоми та способи вивчення суспільних явищ. Сукупність їх створює статистичну методологію.

Теоретичною основою математичної статистики є теорія ймовірностей, а також теорії тих галузей знань, які досліджують відповідні масові процеси та явища.

При використанні статистичного методу розглядається сукупність факторів, які характеризуються певними статистичними ознаками. Це пов'язано з тим, що основним завданням статистики є абстрагування від випадкового та виявлення типового, закономірного. Якщо закономірності природи виявляються в кожному окремому явищі, то статистичні закономірності виявляються тільки в масовій кількості одиниць спостереження, тобто у великих сукупностях. У зв'язку з цим статистичну закономірність можна встановити лише при масовому спостереженні завдяки дії закону великих чисел.

Зміст положень закону великих чисел полягає в тому, що із збільшенням кількості спостережень вплив випадкових причин, що

визначають величину ознаки окремих одиниць сукупності, в цілому взаємно погашається і у зведених характеристиках виражається дія основних причин, тобто визначається закономірність.

Звичайно, закон великих чисел не є регулятором масових процесів і явищ. Велика кількість одиниць сукупності є тільки необхідною умовою для чіткого прояву внутрішнього закону явища. Наприклад, факт народження у матері хлопчика або дівчинки ніскільки не задежить від дії закону великих чисел. Проте, спираючись на масові спостереження, статистика встановлює закономірність, згідно з якою на кожні 100 дівчаток народжується 106 хлопчиків. При цьому у матерів у віці 18 — 20 років на кожні 100 дівчаток народжується 120 хлопчиків, а у віці 38 — 40 років — лише 90.

Закон великих чисел відіграє важливу роль у статистиці. Свідомо використовуючи його, статистика виявляє закономірності масових явищ.

Метод статистики грунтується на поєднанні аналізу та синтезу. З одного боку, статистика виділяє й окремо досліджує складові частини явища, які різняться умовами формування та стадіями розвитку, а з другого, — узагальнюючи дані по окремих частинах, відображує тенденції і форми розвитку явища в цілому.

Будь-яке статистичне дослідження складається з трьох стадій:

1) статистичного спостереження;
2) систематизації матеріалів статистичного спостереження;
3) аналізу узагальнених матеріалів, добутих у результаті їх обробки.

Статистичне спостереження — перша стадія статистичного дослідження. На цій стадії завданням статистики є облік кожної одиниці сукупності та індивідуальних значень властивих їй ознак. Характерним для цієї стадії є метод масового спостереження. Тільки масове спостереження дає змогу виявитися загальним умовам, які характерні для всієї сукупності, і уникнути впливу випадкових причин, що діють на окремі елементи сукупності.

Систематизація матеріалів статистичного спостереження — друга стадія статистичного дослідження. На цій стадії зібрані факти

класифікуються та систематизуються. Результати статистичного спостереження поділяються за ознаками й об'єднуються в однородні групи у вигляді рядів розподілу.

Рядом розподілу називають розподіл одиниць сукупності по групах за величиною варіюючої ознаки.

На другій стадії статистичного дослідження переходять від характеристики одиничних фактів до характеристики їх сукупності, від вивчення індивідуальних значень ознаки до її узагальнення у всій масі одиниць спостереження.

Аналіз зведеного й опрацьованого матеріалу, виявлення закономірностей і зв'язків явищ — третя стадія статистичного дослідження. Систематизовані дані статистичного спостереження підлягають дальшому опрацюванню й аналізу для визначення закономірностей і зв'язків у фактах, що вивчаються, і характеристики типових їх ознак. На цій стадії застосовують специфічний для статистики метод узагальнених показників (абсолютні, середні та відносні величини, показники варіації, кореляції, регресії тощо).

Математична статистика є складовою частиною статистичної науки взагалі, її інструментарієм і специфічним методом дослідження. Лише користуючись математико-статистичними методами соціально-економічна статистика здатна проникнути в специфічну природу досліджуваних масових явищ і виявити закономірності їх розвитку.

Методи математичної статистики широко використовуються в різних галузях діяльності людей, зокрема в аналізі та прогнозуванні розвитку соціально-економічних процесів, дослідно-експериментальній роботі, охороні здоров'я, банківській справі, торгівлі, при контролі якості продукції, визначенні надійності систем автоматизованого управління і т.д.

Контрольні питання

1. *Що є предметом математичної статистики?*
2. *Які теоретичні основи математичної статистики?*
3. *Яке місце математичної статистики в системі статистичних дисциплін?*
4. *У чому полягає різниця між теорією ймовірностей, математичною статистикою та загальною теорією статистики?*
5. *Історія розвитку математичної статистики.*
6. *Завдання математичної статистики.*
7. *Назвіть галузі застосування методів математичної статистики.*

Розділ 2
СТАТИСТИЧНІ РЯДИ РОЗПОДІЛУ ТА ЇХ ХАРАКТЕРИСТИКИ

2.1 Поняття про ряди розподілу та їх види

Будь-яке статистичне дослідження природно — економічних явищ і процесів починається із збирання відповідної первинної інформації: масових матеріалів, даних і фактів. Цю інформацію дістають за допомогою статистичного спостереження.

Статистичне спостереження — це науково організоване збирання масових даних про явища і процеси за допомогою реєстрації суттєвих ознак за зазделегідь розробленою програмою. Воно є одним з найважливіших методів статистики і першим етапом статистичного дослідження.

Від правильної організації та якісного проведення статистичного спостереження значною мірою залежить успіх усього дослідження. Якщо статистичне спостереження проведено неправильно й одержано невірогідні дані, то, як би добре вони не були опрацьовані, матеріали статистичного дослідження будуть низької якості, оскільки недоліки первинних даних не можна усунути їх дальшою обробкою.

Матеріали статистичного спостереження — це первинна інформація, яка підлягає подальшій систематизації і обробці. Основною формою систематизації первинної статистичної інофрмації є ряд розподілу.

Найпростішим видом ряду розподілу є ранжирований ряд, у якомёу одиниці сукупності розміщені в порядку зростання або спадання варіюючої ознаки. Проте такий ряд не дає загальної картини розподілу, оскільки не дозволяє судити про закономірності, закладені в розподілі даних, не показує навколо якої величини концентрується більшість показників. У зв'язку з цим виникає потреба подальшої обробки статистичної інформації, об'єднання її в окремі групи та підрахунку кількості одиниць, що входять у кожну групу.

Статистичним рядом розподілу називають розподіл одиниць сукупності за групами за величиною варіюючої ознаки. Такі ряди складаються з двох елементів: переліку груп і кількості одиниць, що входять у кожну групу. Вони характеризують склад сукупності за розміром досліджуваної ознаки.

Ряди розподілу можна утворити за якісною (атрибутивною) або кількісною ознакою. Відповідно до цього розрізняють два види рядів — атрибутивні та варіаційні. Прикладом атрибутивного ряду може бути розподіл бірж за типами, які виникли в умовах розвитку багатоукладної економіки та ринкових відносин в Україні (табл. 1).

Таблиця 1 *Кількість бірж в Україні у 2002 р.*

Типи бірж	Кількість
Усього зареєстровано	429
у тому числі: універсальні	113
товарні й товарно — сировинні	176
агропромислові	32
Фондові	27
Інші	81

Різновидом атрибутивних рядів є альтернативні ряди. **Альтернативними** називають ряди якісних ознак, які мають два значення, що виключають одне одного: так або ні. Прикладом таких рядів може бути розподіл аграрних формувань на прибуткові та збиткові, або на такі, що виконали і не виконали планового завдання тощо.

Варіаційний ряд розподілу складається з варіантів і частот. Варіантами називають окремі значення групувальної ознаки, а частотою — кількість одиниць спостереження, що мають однакове значення ознаки. Іноді замість частот використовують частості. **Частість** — це відносна величина структури, тобто частка частоти варіанта в загальній сумі частот. Нагромаджені (кумулятивні) частоти (частості) — це сума частот (частостей) варіантів від мінімального значення ознаки до даного.

Розрізняють дискретні та інтервальні варіаційні ряди. **Дискретними** називають ряди, в яких варіанти виражені цілими числами.

Прикладом такого ряду може бути розподіл домогосподарств Харківської області за кількістю осіб (табл. 2).

Таблиця 2 *Розподіл домогосподарств за кількістю осіб (за даними статистичного обстеження 2002 р.)*

Кількість осіб	Кількість домогосподарств	
	частота, тис.	частість, %
1	218	20,0
2	333	30,6
3	263	24,2
4	189	17,4
5 і більше	86	7
Разом	1089	100 %

Прикладом інтервального варіаційного ряду може бути розподіл 250 корів за живою масою (**інтервальними** називають ряди, в яких варіанти виражені у вигляді інтервалів (табл. 3).

Таблиця 3 *Розподіл 250 корів за живою масою*

Жива маса корови, кг	400 – 420	420 – 440	440 – 460	460 – 480	480 – 500	500 – 520	520 – 540	540 – 560	560 – 580
Поголів'я корів	5	21	35	42	57	40	31	15	4

У наведеному прикладі варіантами є значения маси корови, а частотами — кількість корів.

Інтервали в рядах розподілу можуть бути рівні і нерівні. Прикладом ряду розподілу з нерівними інтервалами може бути розподіл аграрних приватних підприємств регіону за чисельністю працюючих (табл 4).

Таблиця 4 *Розподіл аграрних приватних підприємств за кількістю працівників*

Кількість працівників, чол.	Кількість підприємств	
	одиниць	частість, %
до 100	8	16
101 – 200	24	48
201 – 400	13	26
401 і більше	5	10
Разом	50	100

В інтервальних рядах розподілу з нерівними інтервалами визначають такі математичні характеристики, як щільність розподілу і відносна щільність розподілу. Першу характеристику визначають відношенням частоти до величини інтервалу, другу — відношенням частості до величини того ж інтервалу. Для наведеного в табл. 4 прикладу щільність розподілу на другому інтервалі становить 0,24 (24:100), а відносна щільність на цьому інтервалі 0,48 % (48:100).

Залежно від завдань дослідження ряди розподілу можна будувати за будь — якою ознакою. У практичній статистиці їх широко застосовують для вивчення розподілу аграрних формувань за розміром посівних площ, урожайністю культур, поголів'ям тварин,

продуктивністю праці, рентабельністю продукції тощо. Особливе значення має порівняння двох або кількох рядів розподілу різних сукупностей або тієї самої сукупності за різні періоди. Таке порівняння дає змогу оцінити розбіжності між сукупностями і дослідити структурні зрушення. При різних розмірах сукупностей порівняння краще користуватися не частотами, а частостями.

2.2 Визначення кількості груп і величини інтервалів у рядах розподілу

Особливе значення при використанні рядів розподілу в економіко — статистичному аналізі має визначення кількості груп і величини інтервалів, які показують мінімальне та максимальне значення ознаки для кожної групи.

Як відмічалося раніше, групувальні ознаки можуть бути атрибутивними (якісними) і кількісними. До **атрибутивних** належать такі ознаки, які не мають кількісного виразу та реєструються у вигляді текстового запису (наприклад, групи і різновиди грунтів, породи тварин, марки машин, професії працівників).

Кількісні ознаки реєструються числом. Групувальна ознака може бути виражена числами по-різному. Одні ознаки виражаються тільки цілими числами (наприклад, кількість машин, поголів'я тварин, кількість працівників). Така ознака називається **дискретною,** або **перервною**. Інші ознаки можуть позначатися цілими та дробовими числами. Так, рівень урожайності сільськогосподарських культур можна визначити з точністю до цілого числа, десятих, сотих і т.д. Ці зміни ознаки називають безперервними.

При групуванні за атрибутивною ознакою кількість груп, на які поділяється досліджувана сукупність, визначають кількістю різновидів (градацій) цієї ознаки. Наприклад, при групуванні посівних площ озимої пшениці за сортами груп буде стільки, скільки сортів вирощується в господарстві. Так само визначають кількість груп, якщо групувальна ознака змінюється дискретно і має обмежене число значень (наприклад, кількість дітей у сім'ї). Якщо при цьому окремі групи мають обмежену кількість одиниць і є близькими за змістом, то їх об'єднують у більші якісно однорідні групи.

При групуванні за кількісною ознакою, яка змінюється безперервно (урожайність культур, продуктивність тварин, обсяг і собівартість продукції) або змінюється дискретно, але у широких межах (поголів'я тварин, чисельність працівників і т.д.), кількість груп

залежить від чисельності досліджуваної сукупності і характеру варіювання групувальної ознаки. Виділені групи мають бути якісно однорідними, а в кожній групі повинно бути стільки одиниць, щоб уникнути випадковості неправильних висновки.

Якщо групувальна ознака має плавний характер варіювання і при цьому застосовуються рівні інтервали, то кількість груп орієнтовно можна визначити за формулою американського вченого Стерджеса

$$n = 1 + 3{,}322\ lg\ N,$$

де n — кількість груп; N — чисельність сукупності.

Користуючись цією формулою, можна скласти номограму, яка показує залежність кількості груп від чисельності сукупності:

N	15 – 24	25 – 44	45 – 89	90 – 179	180 – 359	360 – 719	720 – 1439
n	5	6	7	8	9	10	11

Слід урахувати, що наведена вище формула одержана емпіричним шляхом і не враховує ні складу, ні характеру варіювання ознаки досліджуваної сукупності. Тому механічне її застосування може призвести до неправильних результатів.

Визначення кількості груп за наведеною вище формулою дає правильні результати тільки тоді, коли варіювання групувальної ознаки має плавний характер, а сукупність достатньо чисельна. У всіх інших випадках кількість груп слід визначати на підставі аналізу варіювання групувальної ознаки з тим, щоб кількісні зміни між окремими групами відображали якісні зміни досліджуваного явища. Для цього спочатку потрібно утворити ранжирований ряд, у якому одиниці сукупності розподілити у порядку зростання чи зменшення групувальної ознаки. Аналізуючи цей ряд, особливо його графічне зображення, можна виділити якісно відмінні групи.

На основі ранжированого ряду можна побудувати варіаційний ряд розподілу та визначити кількість істотно відмінних однорідних груп.

При групуванні за кількісною ознакою важливим є визначення інтервалу групування. **Інтервалом групування** називається різниця між максимальним і мінімальним значеннями ознаки в кожній групі.

За величиною інтервали поділяють на рівні та нерівні. Якщо варіація групувальної ознаки незначна, а розподіл одиниць сукупності має порівняно рівномірний характер, то застосовують **рівні інтервали**, наприклад, при групуванні посівних площ сільськогосподарських культур за рівнем урожайності, працівників окремих галузей і професій за рівнем заробітної плати тощо.

Величину інтервалу при групуванні із застосуванням рівних інтервалів визначають за формулою

$$i = \frac{x_{max} - x_{min}}{n},$$

де i — величина інтервалу; x_{max} — максимальна величина групувальної ознаки; x_{min} — мінімальна величина групувальної ознаки; n — кількість груп.

Припустімо, що якість грунтів аграрних формувань коливається у межах від 60 до 80 балів. Потрібно згрупувати агроформування за якістю грунтів, утворивши при цьому чотири групи з рівними інтервалами. Величина інтервалу становитиме

$$i = \frac{80 - 60}{4} = 5.$$

Відповідно до цього інтервали групування будуть такими: 60 – 65; 65 – 70; 70 – 75; 75 – 80 балів.

Якщо групувальна ознака змінюється в значних розмірах, застосовують нерівні (прогресивні) інтервали. **Нерівними** називають інтервали, в яких різниця між верхньою та нижньою межею неоднакова.

Так, при вивченні розподілу аграрних формувань за кількістю великої рогатої худоби на перше січня кожного року використовують такі інтервали: 1 – 99 голів; 100 – 299; 300 – 599; 600 – 999; 1000 – 3000, понад 3000.

Інтервали не можна робити великими, оскільки в одну групу можуть потрапити якісно різнорідні одниці. Не можна робити інтервали і дуже малими, тому що в окремих групах буде мало спостережень і висновки можуть бути ненадійними.

При групуванні за кількісною ознакою необхідно правильно позначати нижню та верхню межу кожної групи. Коли ознака вимірюється тільки цілими числами (наприклад, чисельність працівників, поголів'я тварин, кількість машин і т.д.), то межі інтервалів позначаються «від – до». Наприклад, при групуванні сервісних кооперативів за кількістю тракторів межі інтервалів позначають так: від 1 до 5 тракторів; від 6 до 10; від 11 до 15 і т.д.

Якщо групувальна ознака змінюється безперервно, то інтервали позначають так, що одне і те саме число є верхньою та нижньою межею двох суміжних груп. Наприклад, при групуванні сільськогосподарських підприємств за врожайністю зернових культур межі інтервалів позначають так: від 20 до 25 ц; від 25 до 30; від 30 до 35 ц і т.д. При цьому використовують правило, що ліве число включає відповідне значення ознаки, а праве — не включає. Так, підприємство, яке має урожайність 25 ц/га, буде віднесене до другої групи.

У статистичній практиці також застосовують закриті і відкриті інтервали. **Закритими** називають інтервали, в яких відомі мінімальні та максимальні значення ознаки (наприклад, групування трактористів-машиністів за стажем роботи: 0 – 5, 5 – 10, 10 – 15 років і т. д.).

Відкритими називають інтервали, в яких невідомі мінімальні та максимальні значення. Відкритими можуть бути перший і останній інтервали, наприклад при групуванні трактористів-машиністів за кількістю відпрацьованих днів за рік: до 200 днів, 200 – 250, 250 – 300, 300 днів і більше.

2.3 Графічне зображення рядів розподілу

Для того, щоб наочно мати уявлення про характер розподілу, застосовують графічне зображення рядів розподілу. Основними способами графічного зображення рядів розподілу є огіва, гістограма, полігон, кумулята і крива Лоренца. Їх будують у системі прямокутних координат, де на горизонтальній осі абсцис відкладають значення варіантів, а на вертикальній осі ординат — частоти (частості).

Огіва — графічне зображення ранжированого ряду розподілу. На осі абсцис відкладають номер господарства у ранжированому ряду, а на осі ординат — значення досліджуваної ознаки (варіанти). Послідовність побудови огіви покажемо на прикладі даних (табл. 5).

Таблиця 5 *Ранжированый ряд господарств за рівнем урожайності зернових культур*

Номер господарства	Урожайність зернових культур, ц/га	Номер гоподарства	Урожайність зернових культур, ц/га
1	25,1	5	36,4
2	26,2	17	37,5
8	26,8	6	37,5
7	27,5	9	38,0
15	28,4	10	39,5
4	30,6	16	39,8
11	31,8	13	40,4
3	32,4	14	41,3
18	35,1	12	42,5

Дані табл. 5 характеризують ранжированый ряд господарств району, в якому одиниці сукупності розміщені у порядку зростання урожайності зернових культур. Ранжирування є початковим етапом побудови варіаційних рядів розподілу. Його використовують для графічного зображення інтенсивності й швидкості зміни досліджуваної ознаки (рис. 1).

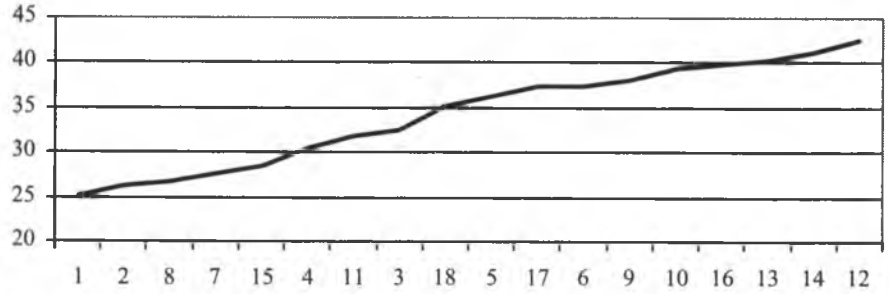


Рис. 1 *Огіва ранжированого ряду розподілу господарств за урожайністю зернових культур*

Найчастіше варіаційні ряди зображують у вигляді **гістограми** (стовпчикової діаграми розподілу) і **полігона** (миогокутника розподілу).

Гістограму застосовують для зображення інтервальних варіаційних рядів. При її побудові на осі абсцис відкладають відрізки, які зображують інтервал. Площа кожного стовпчика повинна бути пропорційною частотам (частостям). Для рівних інтервалів ширину стовпчиків беруть однаковою, а висота має бути пропорційною частотам. При нерівних інтервалах ширина стовпчиків має бути пропорційною величині інтервалу у кожній групі, а висоту стовпчиків зменшують у стільки разів, у скільки збільшується величина інтервалу. Побудуємо гістограму розподілу за даними табл. 3 (рис. 2).

Гістограма наочно характеризує особливості розподілу одиниць окремої сукупності за досліджуваною ознакою. Недоліком гістограм є те, що вони не дають можливості порівнювати кілька розподілів.

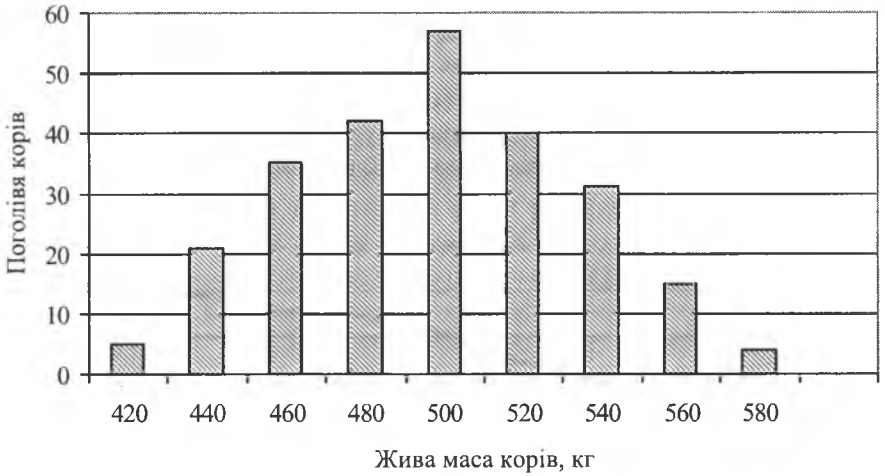


Рис. 2 *Гістограма розподілу корів за живою масою*

Полігон розподілу застосовують для зображення дискретних та інтервальних варіаційних рядів. Координатами точок при цьому є варіанти (в інтервальних рядах середини інтервалів) і частоти, що відповідають цим варіантам. Точки з'єднують прямими лініями. Щоб замкнути полігон, крайні точки з'єднують із серединами інтервалів, у яких частоти дорівнюють нулю (рис. 3).

Полігон розподілу можна дістати з гістограми. Для цього з'єднують прямими лініями середні точки верхніх площадок гістограми. У свою чергу, гістограму можна дістати з полігона, відповідно розмістивши стовпчики.

Кумулята — графічне зображення варіаційного ряду з нагромадженими частотами (частостями). Для її побудови на осі абсцис відкладають варіанти, а на осі ординат — нагромаджені частоти (частості), які показують, скільки одиниць сукупності мають значення ознаки, що не перебільшує цього значення. Кумуляту застосовують при порівнянні різних варіаційних рядів, а також в економічних дослідженнях, зокрема для аналізу концентрації виробництва.

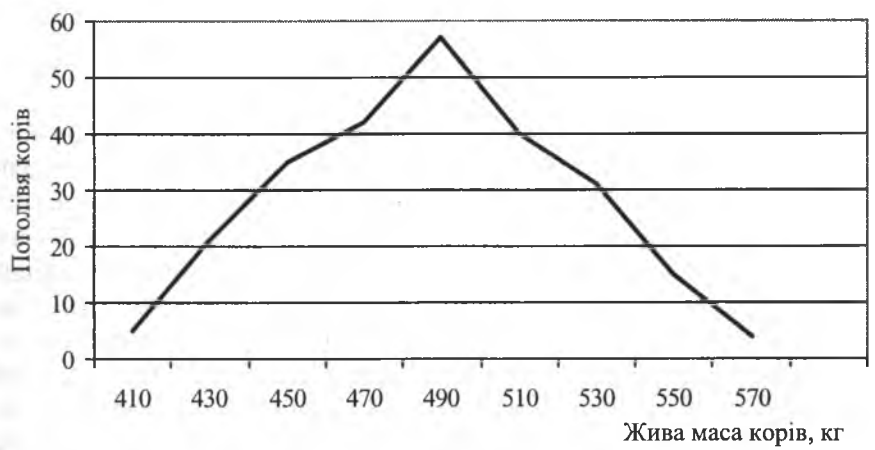


Рис. 3 *Полігон розподілу корів за живою масою*

Побудуємо кумуляту за даними табл. 3 (рис. 4).

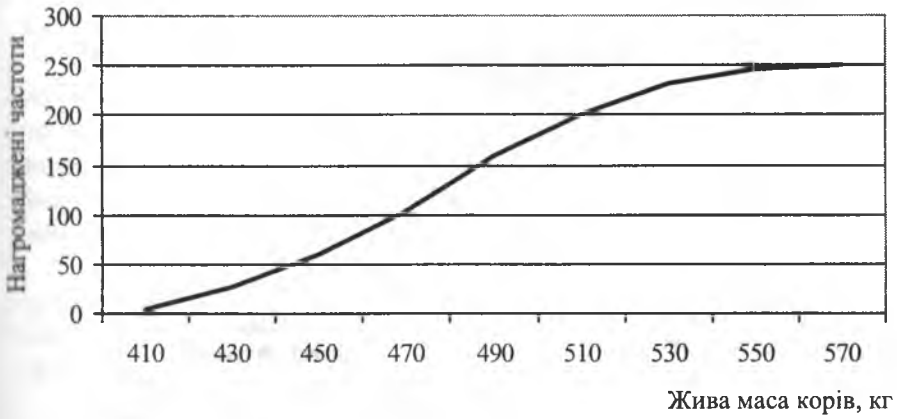


Рис. 4 *Кумулята інтервального розподілу корів за живою масою*

Окремі методи статистики грунтуються на використанні кумуляти. До них належить критерій лямбда, який визначає вірогідність різниць двох розподілів.

Крива Лоренца — це спосіб графічного зображення рівня концентрації явища. Для її побудови на осі координат наносять процентну масштабну шкалу від 0 до 100. На осі абсцис відкладають нагромаджені частості, а на осі ординат — нагромаджені відносні обсяги варіюючої ознаки, вираженої в процентах. Загальний обсяг явища одержують, підсумовуючи добутки варіантів на їх частості. При рівномірному розподілі ознаки певному проценту одиниць сукупності відповідає такий самий процент значень досліджуваної сумарної ознаки. На графіку лінією рівномірного розподілу є діагональ квадрата. При нерівномірному розподілі лінія концентрації становить увігнуту криву залежно від рівня концентрації ознаки. Її відхилення від діагоналі квадрата характеризує ступінь концентрації явища (рис. 5). Чим більша нерівномірність у розподілі частот, тим більше відхилення лінії концентрації від діагоналі.

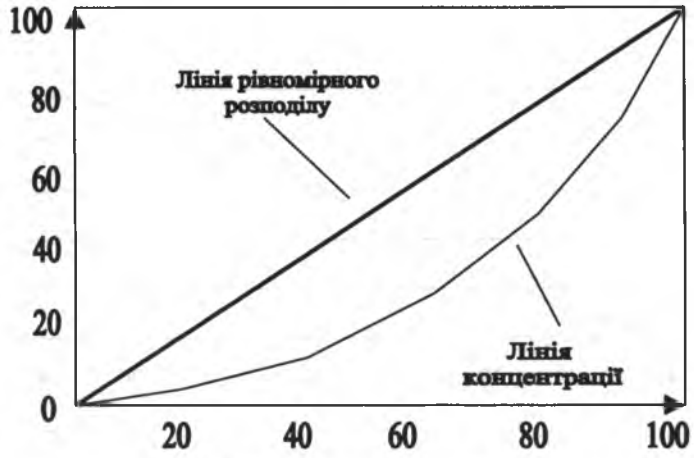


Рис. 5 *Крива Лоренца*

Кожний ряд розподілу можна зобразити за допомогою відповідної кривої, яку називають кривою розподілу. Форми таких кривих залежать від різних факторів, що впливають на формування значень групувальних ознак. Форми графіків розподілу поділяють на два види: одновершинні та багатовершинні.

До **одновершинних** належать такі **криві**, графіки розподілів яких мають один максимум частот. Їх поділяють на **симетричні** (в яких максимум частот припадає на середину ряду) і **асиметричні** (в яких максимум частот припадає на початок або кінець ряду).

Розрізняють правосторонню та лівосторонню асиметрію. Якщо частоти наростають швидше, ніж спадають, асиметрія буде **правосторонньою**, а якщо навпаки — **лівосторонньою**. Причинами асиметрії можуть бути недостатньо велика чисельність сукупності, неправильно проведене групування, а також вплив факторів, які зсувають частоту в той чи інший бік.

До **багатовершинних** належать такі **криві**, графіки розподілів яких мають кілька максимумів частот. Багатовершинність розподілу свідчить про неоднорідність сукупності. У такому разі досліджувану сукупність потрібно розчленувати на окремі однорідні сукупності і вивчати їх окремо.

Розглянуті криві є **фактичними (емпіричними) кривими розподілу**. Вони характеризують залежність між значеннями варіюючої ознаки та відповідними частотами. Фактичні криві розподілу відображають вплив на досліджувану ознаку двох груп факторів: **систематичних**, які є основними причинами, що впливають на характер розподілу, і **випадкових**, які спричинюють відхилення від закономірного розвитку.

Теоретична крива розподілу відображує вплив на варіюючу ознаку лише систематичних факторів. Найбільш поширеними типами теоретичних кривих розподілів є нормальний розподіл, розподіл Фішера, розподіл Стьюдента, розподіл Пірсона тощо.

Порівняння емпіричного і теоретичного розподілів показує, якою мірою досліджувана сукупність виражає закономірності, властиві даному явищу внаслідок впливу основних причин.

2.4 Поняття центральної тенденції розподілу. Середні величини як характеристики ряду

Статистичні величини бувають індивідуальні та середні. Індивідуальні величини виражають розмір ознак окремих одиниць сукупності. Середні величини — це узагальнюючі кількісні показники, що характеризують типові розміри варіюючих ознак якісно однорідних сукупностей.

Індивідуальні значення ознаки в сукупності визначаються як загальними для всіх одиниць причинами, так і особливостями кожної одиниці. Так, продуктивність тварин залежить як від умов годівлі та способів утримання, загальних для ферми або господарства, так і від особливостей окремих тварин і догляду за ними.

Середні величини відображують значення ознаки, характерне для загальних, визначальних умов усієї сукупності. При розрахунку середньої величини індивідуальні коливання ознаки абстрагуються за окремими одиницями сукупності. Унаслідок випадкового характеру та різного напряму ці коливання за сукупністю взаємно урівноважуються, погашаються при усередненні та в середній величині виявляються загальні властивості, характерні для даного масового явища.

Середні величини характеризують центральну тенденцію в розподілі. Центральна тенденція — це властивість значень досліджуваної ознаки групуватися навколо центру розподілу частот, статистичною характеристикою якого є середня величина. Середня величина є узагальнюючою характеристикою центральної тенденції тільки тоді, коли вона обчислена для однорідної сукупності, окремі варіанти якої концентруються поблизу деякого значення. Якщо сукупність неоднорідна, варіанти значно відрізняються один від одного і не виявляють загальної тенденції. Обчислена середня при цьому буде формальною величиною, оскільки вона не є типовою характеристикою сукупності.

Середні величини використовують для узагальненої характеристики сукупностей за істотними ознаками, для порівняння цих ознак у різних сукупностях. Так, на підставі порівняння даних про валові збори зерна по окремих підрозділах господарства не можна зробити правильні висновки про ефективність їх роботи, оскільки розмір посівної площі у підрозділах неоднаковий. Щоб правильно оцінити роботу підрозділів, слід обчислити і порівняти показники, які характеризують середній збір зерна з розрахунку на одиницю сукупності, тобто середню урожайність з 1 га посівної площі.

Середні величини застосовують і при дослідженні закономірностей і тенденцій розвитку суспільних явищ, а також для оцінки ефективності різних форм господарювання, впровадження досягнень науково-технічного прогресу, освоєння передових економічних, агротехнічних і зооветеринарних заходів, для виявлення внутрішньогосподарських резервів.

Основною умовою наукового застосування середніх величин є якісна однорідність сукупності, для якої обчислюють середню. Отже, обчислення та використання середніх величин мають бути тісно пов'язані з методом статистичних групувань, за допомогою якого різнорідні явища розподіляються на якісно однорідні сукупності.

Середні величини для неоднорідних сукупностей спотворюють типові розміри явищ і через це не мають наукового та пізнавального значення.

У зв'язку з наявністю загальних і часткових сукупностей у статистичній практиці загальні середні доповнюються груповими середніми. Загальна середня показує типовий розмір ознаки якісно однорідної сукупності в цілому, а групові середні — розмір окремих її частин із специфічними властивостями.

У свою чергу загальні та групові середні потрібно доповнювати конкретними прикладами, оскільки у відхиленні індивідуального від загального виявляється не тільки випадкове, а й закономірне явище,

яке ще не набуло масового поширення. Наприклад, аналізуючи урожайність озимої пшениці, не можна обмежитися визначенням тільки загальної середньої врожайності в цілому по господарству. Її треба доповнити показниками врожайності перспективних сортів, орендних підрозділів тощо.

Середні величини показують типовий стійкий розмір ознаки для всіх одиниць сукупності або досить великої їх кількості. Тільки при такій умові можливі випадкові коливання взаємно погашаються і в середній виявляються загальні властивості, характерні для всієї сукупності. При несуцільному обліку мінімально потрібну кількість одиниць сукупності для одержання вірогідних середніх визначають, виходячи з особливостей варіювання усередненої ознаки. Чим сильніше коливання ознаки, в окремих одиниць сукупності, тим більше їх слід узяти для одержання вірогідної середньої.

Середні величини можуть бути абсолютними або відносними. Їх виражають у тих самих одиницях вимірювання, що й усереднені ознаки.

У статистиці застосовують різні види середніх величин: середню арифметичну, середню гармонічну, середню геометричну, середню квадратичну, середню кубічну та ін. Вибір певного виду середньої величини залежить від характеру вихідних даних. Правильну характеристику сукупності за варіюючою ознакою у кожному випадку дає тільки певний вид середньої. Загальною умовою правильного обчислення всіх видів середніх є збереження обсягу варіюючої ознаки при заміні індивідуальних значень ознаки їх середньою.

Середня арифметична є найбільш поширеним видом середніх величин. Її застосовують тоді, коли загальний обсяг варіюючої ознаки для всієї сукупності становить суму індивідуальних значень усередненої ознаки. Наприклад, загальний збір урожаю цукрових буряків є сумою врожаїв цієї культури з кожного гектара

площі, загальний надій молока є сумою надоїв молока від кожної корови, загальний фонд заробітної плати є сумою виплат окремим працівникам тощо.

Середню арифметичну визначають як відношення суми окремих значень ознаки до кількості одиниць сукупності. Розрізняють середню арифметичну просту і зважену.

Середню арифметичну просту застосовують тоді, коли відомі індивідуальні значення усередненої ознаки у кожної одиниці сукупності. Її визначають за формулою:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x}{n},$$

де $\bar{x}$ — середнє значення ознаки; x – окремі значення ознаки (варіанти); n — кількість варіантів.

Середню арифметичну зважену обчислюють тоді, коли окремі значення усередненої ознаки повторюються в досліджуваній сукупності неоднакове число разів, а також для обчислення середньої із середніх при різному обсязі сукупності. Зважування в цьому разі проводять за частотами, які показують, скільки разів повторюється певний варіант. Середню арифметичну зважену визначають за такою формулою:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum xf}{\sum f},$$

де f — частоти (ваги).

Послідовність розрахунку середньої арифметичної простої і зваженої розглянемо на прикладі обчислення середнього розміру посівної площі та середньої урожайності цукрових буряків для групи господарств (табл. 6).

Таблиця 6 *Посівна площа і урожайність цукрових буряків по групі господарств*

Номер господарства	Посівна площа, га	Урожайність, ц/га	Валовий збір, ц
	f	X	xf
1	250	340	85 000
2	400	300	120 000
3	380	320	121 600
4	260	360	93 600
5	310	280	96 800
Разом	1 600	X	507 000

Щоб визначити середній розмір посівної площі цукрових буряків з розрахунку на одне господарство, слід підсумувати посівну площу п'яти господарств і одержану суму поділити на їх кількість:

$$\frac{\sum f}{n} = \frac{250 + 400 + 380 + 260 + 310}{5} = 320 \ га.$$

Ці розрахунки виконані за формулою середньої арифметичної простої, оскільки окремі варіанти (розмір посівної площі кожного господарства) повторюються однакове число разів.

Для визначення середньої урожайності цукрових буряків потрібно їх загальний валовий збір по групі господарств поділити на загальний розмір посівної площі:

$$\bar{x} = \frac{\sum fx}{\sum f} = 507000 : 1600 = 317 ц / га.$$

У цьому разі використано формулу середньої арифметичної зваженої, оскільки окремі значення урожайності цукрових буряків по господарствах (варіанти) мають різну повторюваність.

В інтервальних варіаційних рядах значення усередненої ознаки обчислюють не за конкретними числами, а за величинами, вираженими у вигляді інтервалів («від — до»). Порядок розрахунку середньої в інтервальних рядах розподілу розглянемо на такому прикладі (табл. 7).

Таблиця 7 *Розрахунок середньої арифметичної в інтервальному варіаційному ряду*

Вихідні дані		Розрахункові дані	
Річний надій молока від корови, кг	Кількість корів	Середина інтервалу	Валовий надій молока, кг
	f	X	xf
До 2000	18	1750	31500
2000 – 2500	42	2250	94500
2500 – 3000	80	2750	220000
3000 – 3500	144	3250	468000
3500 – 4000	70	3750	262000
4000 – 4500	32	4250	136000
4500 і більше	14	4750	66500
Разом	400	X	1 279 000

Спочатку слід перейти від інтервальних значень до конкретних величин ознаки. Для цього використовують усереднене значення інтервалу. Його визначають підсумуванням нижньої та верхньої меж і діленням знайденої суми на два. При цьому враховують,

що всередині кожного інтервалу при великій кількості одиниць сукупності значення ознаки розподіляється більш — менш рівномірно. Усереднене значення першого останнього інтервалів, які не мають й чітко позначених меж, обчислюють за величиною наступного або попереднього інтервалу. Щоб знайти усереднене значення першого відкритого інтервалу, від його верхньої межі віднімають половину величини наступного інтервалу. Усереднене значення останнього відкритого інтервалу визначають додаванням до його нижньої межі половини величини попереднього інтервалу.

Використовуючи усереднені значення варіантів, середньорічний надій молока від корови визначають за формулою середньої арифметичної зваженої:

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1279000}{400} = 3198 \ кг.$$

Середня арифметична має деякі математичні властивості, знання та використання яких спрощує техніку її обчислення, зокрема:

1) алгебраїчна сума відхилень варіантів від середньої арифметичної, зваженої відповідними частотами, дорівнює нулю:

$$\sum(x-\bar{x}) = \sum x - n\bar{x} = 0 \text{ — для середньої простої;}$$
$$\sum(x-\bar{x})f = \sum xf - \bar{x}\sum f = 0 \text{ — для середньої зваженої.}$$

Ця властивість випливає з положення, що в середній величині взаємно погашаються відхилення від неї окремих варіантів у бік збільшення або зменшення;

2) якщо всі варіанти збільшити (зменшити) на те саме число a, то середня збільшиться (зменшиться) на це ж число a. Збільшимо кожний варіант на постійне число а й обчислимо середню:

$$\frac{\sum(x+a)f}{\sum f} = \frac{\sum xf}{\sum f} + \frac{\sum af}{\sum f} = \bar{x} + \frac{a\sum f}{\sum f} = \bar{x} + a,$$

отже, середня збільшилась на постійне число a;

3) якщо всі варіанти помножити або поділити на те саме число (h) разів, то середня арифметична відповідно збільщиться (зменшиться) у стільки ж разів. Розділимо кожний варіант на постіне число h і визначимо середню:

$$\frac{\sum \frac{x}{h} f}{\sum f} = \frac{1}{h} \bullet \frac{\sum xf}{\sum f} = \frac{\bar{x}}{h},$$

отже, середня зменшилася в h разів;

4) якщо частоти (ваги) окремих варіантів помножити або поділити на яке-небудь число, то значення середньої арифметичної не зміниться. Помножимо частоти окремих варіантів на постійне число а і обчислимо середню:

$$\bar{x} = \frac{\sum xaf}{\sum af} = \frac{a\sum xf}{a\sum f} = \frac{\sum xf}{\sum f} = \bar{x},$$

Отже, значення середньої не змінилося. Залежно від цієї властивості при розрахунку середньої арифметичної з варіантів, частоти яких однакові, замість зваженої середньої можна визначити просту середню, що значно спрощує розрахунки. При цьому варіанти можна зважувати не тільки за абсолютними значеннями частот, а й за їх часткою в загальній сумі.

Для обчислення середньої з варіаційного ряду з рівними інтервалами використовують спосіб відліку від умовного нуля, який ще називають способом моментів. При цьому одне із значень варіюючої ознаки, яке знаходиться у центрі розподілу або має найбільшу частоту, беруть за початок відліку. Потім віж кожного усередненого значення інтервалу віднімають величину, взяту за початок відліку, і різницю ділять на величину інтервалу h. Добуті внаслідок цих перетворень нові варіанти утворюють у варіаційному ряду з рівними інтервалами прості числа — відхилення в інтервалах від умовного нуля.

Щоб обчислити середню арифметичну, середнє відхилення в інтервалах множать на величину інтервалу і додають величину, взяту за початок відліку:

$$\bar{x} = \frac{\sum\left(\frac{x-a}{h}\right)f}{\sum f}h + a,$$

де a – величина, взята за початок відліку; h – величина інтервалу.

Розглянемо визначення середньої арифметичної для інтервального ряду способом відліку від умовного нуля (табл. 8).

Таблиця 8 *Розрахунок середньої арифметичної для інтервального ряду способом відліку від умовного нуля*

Надій молока від корови, кг	Кількість корів	Середина інтервалу	Відхилення від умовного нуля	Відхилення в інтервалах	Добуток від множення відхилення в інтервалах на частоти
	f	x	$x-a$	$\frac{x-a}{h}$	$\left(\frac{x-a}{h}\right)f$
До 2000	18	1750	–1500	–3	–54
2000 – 2500	42	2250	–1000	–2	–84
2500 – 3000	80	2750	–500	–1	–80
3000 – 3500	144	3250	0	0	0
3500 – 4000	70	3750	500	1	70
4000 – 4500	32	4250	1000	2	64
4500 і більше	14	4750	1500	3	42
Разом	400	X	X	X	–42

За початок відліку (a) взято усереднений варіант з найбільшою частотою ($a = 3250$). Величина інтервалу $h = 500$. Середній річний надій молока від однієї корови становитиме $\frac{-42}{400} \cdot 500 + 3250 = 3198 кг$.

Середню гармонічну використовують для узагальненої характеристики ознаки тоді, коли відомі окремі значення досліджуваної ознаки й обсяги явищ, а частоти невідомі. Середня гармонічна — це обернена величина середньої арифметичної, обчисленої з обернених значень усереднених ознак. Вона буває простою і зваженою. Формула середньої гармонічної простої має такий вигляд:

$$\bar{x}_r = \frac{n}{\sum\frac{1}{x}},$$

де x – варіанти, n – кількість варіантів.

Припустимо, що два трактористи — машиністи працюють на сівбі зернових культур. Перший тракторист на сівбу 1 га зернових культур витрачає 20 хв, другий — 30 хв. Потрібно визначити середні затрати праці на сівбу 1 га зернових культур. Розрахунок здійснюємо за формулою середньої гармонічної простої:

$$\bar{x}_r = \frac{n}{\sum\frac{1}{n}} = \frac{2}{\frac{1}{20}+\frac{1}{30}} = \frac{2}{\frac{5}{60}} = 24 \text{ хв.}$$

Щоб переконатися у правильності розрахунку, загальні затрати робочого часу двома трактористами — машиністами за робочу зміну (7·60·2 = 840 хв.) ділять на загальну засіяну площу (420:20+420:30 = = 35 га). Використання у цьому разі середньої арифметичної простої [(20=30):2 = 25 хв.] дає неправильний результат.

На практиці частіше застосовують середню гармонічну зважену, формула якої має такий вигляд:

$$\bar{x}_r = \frac{\sum \omega}{\sum \frac{\omega}{x}},$$

де ω — обсяги явищ.

Розглянемо порядок розрахунку середньої гармонічної зваженої на такому прикладі (табл. 9).

Таблиця 9 *Розрахунок середньої гармонічної зваженої*

Культура	Урожайність, ц/га	Валовий збір, ц	Частота
	f	ω	$\frac{\omega}{x}$
Озима пшениця	45	22 500	500
Ячмінь	36	7 200	200
Кукурудза	50	15 000	300
Гречка	18	900	50
Разом	X	45 600	1050

Оскільки середня врожайність зернових культур становить відношення загального валового збору зерна до загальної посівної площі, то спочатку визначимо розмір посівної площі по кожній культурі, а потім — середню величину:

$$\bar{x}_r = \frac{\sum \omega}{\sum \frac{\omega}{x}} = \frac{45600}{1050} = 43{,}4 \ \ ц/га.$$

Середня гармонічна, як і середня арифметична, не зміниться, якщо обсяги явищ, які є вагами окремих варіантів, помножити (поділити)

на яке-небудь число. Це дає змогу при її обчисленні користуватися не абсолютними показниками, а їх частками. Припустимо, потрібно визначити середню ціну реалізації 1 т молока (табл. 10).

Таблиця 10 *Розрахунок середньої ціни реалізації молока*

Сорт молока	Ціна реалізації 1 т, грн.	Частота виручки у загальній сумі, %	Відношення частоти виручки до ціни реалізації
	x	d	$\frac{d}{x}$
1	554	78	0,141
2	507	17	0,034
Несортове	470	5	0,011
Разом	X	100	0,186

У нашому прикладі немає виручки від реалізації окремих сортів молока, яка становить добуток ціни реалізації на кількість проданого молока. Тому замість обсягів явищ можна використати їх співвідношення, тобто частку окремих партій (сортів) молока у загальній сумі виручки. Звідси матимемо:

$$\bar{x} = \frac{\sum d}{\sum \frac{d}{x}} = \frac{100}{0{,}186} = 538 \ \ грн.$$

Середню гармонічну зважену застосовують також для обчислення середньої урожайності по групі однорідних культур або групі господарств, якщо відомі валовий збір і урожайність окремих культур або в господарствах; для визначення середнього процента

виконання плану виробництва і реалізації продукції по однорідній сукупності, якщо відомі дані про кількість фактично виробленої продукції та рівень виконання плану по окремих об'єктах тощо.

Середню геометричну використовують для визначення середніх темпів зростання, тобто коли загальний обсяг явищ становить не суму, а добуток ознак x. Наприклад, виробництво продукції в господарстві за перший рік збільшилося в 1,12 раза, за другий — в 1,09, за третій — 1,10, за четвертий — 1,12, за п'ятий — 1,14, за шостий рік — в 1,13 раза. Потрібно визначити середнє збільшення обсягу виробництва продукції за рік. Загальний обсяг виробництва продукції у цьому разі дорівнюватиме добутку щорічних коефіцієнтів зростання: 1,12 · 1,09 ·1,10 ·1,12 · 1,14 x x 1,13 = 1,94. Отже, обсяг виробництва продукції в останньому році збільшився порівняно з базисним роком у 1,94 раза. Кількість варіантів $n = 6$.

Середній темп зростання за рік обчислимо за формулою середньої геометричної

$$\bar{x}_{геом} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n} = \sqrt[n]{Д_x},$$

де $Д$ – добуток.

Корінь ступеня n з добутку варіантів визначимо логарифмуванням:

$$\lg \bar{x}_{геом.} = \frac{1}{n}(\lg x_1 + \lg x_2 + \lg x_3 + \ldots + \lg x_n).$$

У нашому прикладі:

$$\lg \bar{x}_{геом.} = \frac{1}{6}(\lg 1{,}12 + \lg 1{,}09 + \lg 1{,}10 + \lg 1{,}12 + \lg 1{,}14 + \lg 1{,}13) = 0{,}0479.$$

Користуючись таблицею антилогарифмів, визначимо $\bar{x}_{геом.} = 1{,}117$. Отже, виробництво продукції за рік збільшилось в середньому в 1,117 раза, або становить 111,7 %.

Середню геометричну просту застосовують тоді, коли періоди, за які визначені усереднені коефіцієнти зростання, дорівнюють один одному.

Якщо ж відомі середні коефіцієнти зростання за неоднакові за тривалістю періоди і слід визначити середній коефіцієнт зростання за весь період, використовують середню геометричну зважену:

$$\bar{x}_{геом.} = \sum f\sqrt{\bar{x}_1^{f_1} \cdot \bar{x}_2^{f_2} \ldots \cdot \bar{x}_n^{f_n}},$$

де $\bar{x}_{геом.}$ — середній темп зростання за весь період; $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_n$ — середні темпи зростання за окремі періоди; f — тривалість окремих періодів (ваги).

Порядок розрахунку середньої геометричної зваженої розглянемо за даними нашого прикладу. Середні коефіцієнти зростання виробництва продукції дорівнюють:
за перші два роки

$$\bar{x}_{геом.} = \sqrt[2]{1{,}12 \cdot 1{,}09} = 1{,}105;$$

за останні чотири роки

$$\bar{x}_{геом.} = \sqrt[4]{1{,}10 \cdot 1{,}12 \cdot 1{,}14 \cdot 1{,}13} = 1{,}122;$$

за весь період

$$\bar{x}_{геом.} = \sum n\sqrt{\bar{x}_1^{n_1} \cdot \bar{x}_2^{n_2} \ldots \cdot \bar{x}_n^{n_n}} = \sqrt[2+4]{1{,}105^2 + 1{,}122^4} = 1{,}177.$$

Середню квадратичну використовують в основному для оцінки варіації (мінливості) ознак, а також для узагальнення ознак, виражених лінійними розмірами яких-небудь площ (для розрахунку середніх

діаметрів стовбурів дерев, листків, кошиків тощо). Її визначають за такими формулами:

$$\bar{x}_{кв} = \sqrt{\frac{\sum x^2}{n}} - проста;$$

$$\bar{x}_{кв} = \sqrt{\frac{\sum x^2 f}{\sum f}} - зважена;$$

де x — окремі значення ознаки (варіанти); f — частоти (ваги).

Наприклад, діаметр трьох кошиків соняшнику становить 12, 19 і 23 см. Потрібно визначити середній діаметр кошика.

Насамперед переконаємося, що розрахунок середнього діаметра кошика за формулою середньої арифметичної не дає правильної відповіді. Про це можна судити після визначення середньої площі кошика

$$\bar{x} = \frac{\sum x}{n} = \frac{12+19+23}{3} = 18см,$$

звідки середня площа кошика, обчислена за формулою площі круга, $^1/_4\pi d^2 = {}^1/_4 \cdot 3{,}14 \cdot 182 = 253{,}4$ см^2.

Середня площа кошика, обчислена через індивідуальні значення площі окремих кошиків, становитиме (113,0 + 283,4 + 415,3):3 = = 270,6 см^2.

Правильну відповідь про середній діаметр кошика дає розрахунок за формулою середньої квадратичної:

$$\bar{x}_{кв} = \sqrt{\frac{\sum x^2}{n}} = \frac{\sqrt{12^2+19^2+23^2}}{3} = 18{,}57см.$$

Щоб переконатися у правильності цього розрахунку, підставимо визначений середній діаметр кошика у формулу площі круга: $^1/_4\pi d^2 = {}^1/_4 \cdot 3{,}14 \cdot 18{,}572 = 270{,}6$ см^2.

Якби у наведеному прикладі окремі значення діаметра кошика повторювалися неоднакову кількість разів, то середній діаметр слід було б розраховувати за формулою середньої квадратичної зваженої (табл. 11).

Середній діаметр кошика соняшника:

$$\bar{x}_{кв} = \sqrt{\frac{\sum x^2 f}{\sum f}} = \sqrt{\frac{8851}{27}} = 18{,}1 \ см.$$

Таблиця 11 *Розрахунок середньої квадратичної зваженої*

Діаметр кошика, см	Кількість кошиків	Квадрат діаметра кошика	Добуток квадрата діаметра на частоту
x	f	$x2$	x^2f
12	8	144	1152
19	14	361	5054
23	5	529	2645
Разом	27	X	8851

Середню кубічну використовують для узагальнення ознак, виражених лінійними розмірами об'ємних фігур, наприклад, для розрахунку середніх діаметрів плодів, насіння, пилкових зернин, яєць тощо. Її обчислюють за такими формулами:

$$\bar{x}_{куб.} = \sqrt[3]{\frac{\sum x^3}{n}} - проста;$$

$$\bar{x}_{куб.} = \sqrt[3]{\frac{\sum x^3 f}{\sum f}} - зважена\ .$$

Припустимо, що діаметри плодів яблуні становлять 4, 6, 6, 7, і 7 см. Потрібно визначити середній діаметр плодів. Підставивши ці дані у формулу, матимемо:

$$\bar{x}_{куб.} = \sqrt[3]{\frac{\sum x^3}{n}} = \sqrt{\frac{4^3+6^3+6^3+7^3+7^3}{5}} = 6{,}18\ см;$$

Середній діаметр, визначений за формулою середньої арифметичної, становить 6 см. Середній об'єм плода, обчислений за формулою $^1/_6\ \pi d^3$, при діаметрі 6,18 см дорівнюватиме 123,7 см^3, а при діаметрі 6 см – 113 см^3. Звідси незначна розбіжність у розмірі діаметра призводить до досить істотних розбіжностей в об'ємі плодів. Обчислюючи середній об'єм плода через індивідуальні значення об'ємів, можна переконатися, що у цьому разі середня кубічна дає більш правильну характеристику досліджуваної ознаки.

Таблиця 12 *Розрахунок середньої кубічної зваженої*

Діаметр площа, см	Кількість плодів	Куб діаметра плода	Добуток куба діаметра плода на частоту
x	f	x^3	x^3f
4	9	64	576
5	23	125	2875
6	35	216	7560
8	7	512	3584
Разом	100	X	23513

Порядок розрахунку середньої кубічної зваженої розглянемо на такому прикладі (табл. 12).

Середній діаметр плода яблуні

$$\bar{x}_{куб} = \sqrt[3]{\frac{\sum x^3 f}{\sum f}} = \sqrt{\frac{23513}{100}} = 6{,}2\ см;$$

Особливим видом середніх величин є середня багатовимірна. Вона становить похідну величину, розраховану для статистичної сукупності чисельністю N одиниць з порядковими номерами i (i = 1, 2, 3,…,N), які мають K ознак (x) з порядковими номерами j (j = 1, 2, 3,…,K). Визначаючи середню багатовимірну, спочатку обчислюють відношення рij кожної ознаки (x) окремих одиниць сукупності до її середнього значення

$$p_{ij} = \frac{x_{ij}}{\bar{x}_j},$$

де x_{ij} — значення j-ї ознаки у - ї одиниці сукупності; $\bar{x}_i$ — середнє значення ознаки.

Після цього визначають середню з цих відношень для кожної одиниці сукупності, яка і є багатовимірною середньою:

$$\bar{p}_{i_j} = \frac{\sum p_{i_j}}{K}.$$

Розрахунок середньої багатовимірної розглянемо на такому прикладі (табл. 13).

Таблиця 13 *Розрахунок середньої багатовимірної*

Номер господарства	Родючість землі		Частка технічних культур у загальній структурі посівних площ		Валова продукція з розрахунку на 100 га с. — г. угідь		$\sum p_{ij}$	$\bar{p}_{ij} = \frac{\sum p_{ij}}{3}$
	балів x_1	$p_{i_1} = \frac{x_{i_1}}{\bar{x}_1}$	% x_2	$p_{i_2} = \frac{x_{i_2}}{\bar{x}_2}$	тис.грн. x_3	$p_{i_3} = \frac{x_{i_3}}{\bar{x}_3}$		
1	60	0,85	10	0,82	62	0,80	2,47	0,82
2	64	0,90	9	0,74	64	0,83	2,47	0,82
3	65	0,92	10	0,82	65	0,84	2,58	0,86
4	68	0,96	11	0,90	68	0,88	2,74	0,91
5	70	0,99	12	0,98	70	0,90	2,87	0,96
6	70	0,99	12	0,98	70	0,90	2,87	0,96
7	72	1,01	14	1,15	85	1,10	3,26	1,09
8	74	1,04	13	1,07	86	1,11	3,22	1,07
9	75	1,06	14	1,15	90	1,16	3,37	1,12
10	76	1,07	12	0,98	85	1,10	3,15	1,05
11	76	1,07	14	1,15	90	1,16	3,38	1,13
12	80	1,14	15	1,26	94	1,22	3,62	1,21
Разом	—	12,00	—	12,00	—	12,00	36,00	12,00

Середню багатовимірну використовують для аналізу господарської діяльності підприємств агропромислового комплексу, особливо при визначенні ефективності використання виробничого потенціалу сільського господарства.

Ознаки, які мають певні одиниці сукупності, а інші не мають (прибуткове чи збиткове господарство, підприємства, що виконали або не виконали плану тощо), називають альтернативними. При альтернативній ознаці, коли є лише два варіанти, один з яких виключає інший, наявність ознаки позначають через одиницю, а її відсутність — через нуль.

Позначимо частку одиниць сукупності, які мають певну ознаку, через p, а частку одиниць, які не мають ознаки, — через q. Звідси $p + q = 1$, а $1 - p = q$.

Середнє значення альтернативної ознаки, обчислене за формулою середньої арифметичної:

$$\bar{x}_{pq} = \frac{1 \cdot p + 0 \cdot q}{p + q} = \frac{p}{p + q} = p.$$

Отже, середнє значення альтернативної ознаки дорівнює частці одиниць сукупності, які мають певну ознаку. Вона і є узагальненою характеристикою сукупності варіюючої ознаки.

Поряд із середніми величинами типовими характеристиками варіюючих ознак є мода і медіана. Моду та медіану називають структурними або розподільними середніми, оскільки вони характеризують особливості розподілу одиниць сукупності за розміром досліджуваної ознаки.

Модою називають значення ознаки, яке найчастіше повторюється в досліджуваній сукупності. Тобто, це варіант, який має найбільшу частоту. У дискретних варіаційних рядах (значення ознаки виражені цілими числами) модою є варіант, який найчастіше повторюється. Розглянемо обчислення моди для дискретного варіаційного ряду на такому прикладі (табл. 14).

Таблиця 14 *Розподіл 100 сімей за кількістю членів сім'ї*

Кількість членів сім'ї, чол., x	2	3	4	5	6	7
Кількість сімей, f	8	14	37	23	10	8

У нашому прикладі модальною величиною є сім'я з 4 чоловік, оскільки вказаний варіант у цій сукупності сімей має найбільшу частоту — 37 випадків.

Якщо в ряду розподілу всі варіанти повторюються однаково часто, тобто частоти окремих варіантів дорівнюють одна одній, то в цьому разі моду не можна визначити, оскільки її немає. Коли ж у ряду розподілу два варіанти мають найбільші й однакові частоти, то такий ряд має дві моди, а розподіл називають бімодальним.

Бімодальний розподіл указує на якісну неоднорідність сукупності за досліджуваною ознакою.

В інтервальному варіаційному ряду моду визначають за допомогою додаткових розрахунків. Спочатку знаходять модальний інтервал, тобто інтервал, який має найбільшу частоту. Після цього модальне значення ознаки визначають за формулою

$$x_{мо} = x_0 + h \cdot \frac{f_{мо} - f_{мо-1}}{2f_{мо} - f_{мо+1} - f_{мо-1}},$$

де $x_{мо}$ — мода; x_0 — нижня межа модального інтервалу; h — величина модального інтервалу; $f_{мо}$ – частота модального інтервалу; $f_{мо} - 1$ — частота інтервалу перед модальним; $f_{мо} + 1$ — частота інтервалу після модального.

Розрахунок моди в інтервальному варіаційному ряду розглянемо на такому прикладі (табл. 15).

Таблиця 15 *Розподіл 200 корів за живою масою*

Жива маса корови, кг	400—420	420—440	440—460	460—480	480—500	500—520	520—540
Кількість корів	6	19	38	63	44	26	5

Модальне значення живої маси однієї корови

$$x_{мо} = 460 + 20 \cdot \frac{63 - 38}{2 \cdot 63 - 44 - 38} = 471{,}4 \ кг.$$

Отже, в досліджуваній сукупності найчастіше трапляються корови з живою масою приблизно 471 кг.

Медіаною називають значення ознаки в одиниці сукупності, що займає середнє положення в ранжированому ряду розподілу. Вона є центром розподілу сукупності і ділить її на дві рівні за кількістю частини.

Для визначення медіани, особливо у великих за чисельністю сукупностях, користуються рядом нагромаджених частот. При цьому, щоб знайти медіану в сукупності, яка складається з парної кількості одиниць, нагромаджені частоти слід поділити на два і розрахувати середню з двох центральних варіантів. При непарній кількості одиниць сукупності центр розподілу визначають додаванням до суми частот одиниці і діленням знайдених даних на два.

У варіаційному ряду, який характеризує розподіл 100 сімей за кількістю членів сім'ї (див. табл. 14), медіаною є сім'я з чотирьох чоловік, оскільки цей варіант лежить в центрі розподілу (100:2 = 50) даної сукупності.

Щоб визначити медіану в інтервальному варіаційному ряду, спочатку знаходять медіанний інтервал, тобто інтервал, у якому є значення ознаки, що ділить ряд розподілу на дві рівні частини. Після цього медіанне значення ознаки обчислюють за формулою:

$$x_{ме} = x_0 + h\frac{\frac{\sum f}{2} - S}{f_{ме}},$$

де $x_{ме}$ — медіана; x_0 — нижня межа медіанного інтервалу; h — величина медіанного інтервалу; f — частоти; S — сума нагромаджених частот перед медіанним інтервалом; $f_{ме}$ — частота медіанного інтервалу. У нашому прикладі медіанне значення живої маси однієї корови (див. табл. 15)

$$x_{ме} = 460 + 20\frac{\frac{200}{2} - 62}{63} = 472{,}1 \ кг.$$

Отже, половина всіх корів має масу менш як 472,1 кг, а решта — більш як 472,1 кг.

Медіану і моду частіше використовують для характеристики явищ і процесів, до яких немає суцільних даних. У сільськогосподарськнх підприємствах структурні середні застосовують при визначенні типових норм виробітку, коли за базу для розрахунку норм виробітку беруться не середні затрати робочого часу на виконання окремих технологічних операцій, а їх модальне або медіанне значення.

Додатковими статистичними характеристиками рядів розподілу є квартилі та децилі. Квартилі поділяють ряд розподілу за сумою частот на 4 рівні частини, а децилі – на 10 рівних частин. Другий квартиль (Q_2) дорівнює медіані. Перший (Q_1) і третій (Q_3) квартилі обчислюють так само, як і медіану, тільки замість медіанного інтервалу беруть для першого квартиля інтервал, в якому знаходиться варіант, що відокремлює ¼ кількості частот, а для третього – інтервал, в якому знаходиться варіант, що відокремлює ¾ кількості частот.

Визначення першого і третього квартилів розглянемо на прикладі даних табл. 15. Для першого квартиля ¼ частина всіх частот

дорівнює 200:4 = 50. П'ятдесятий варіант знаходиться в інтервалі 440 – 460. Нижня межа його $x_Q = 440$, величина $h = 20$, частота $f_Q = 38$, а сума нагромаджених частот інтервалів, що передують відповідним квартильним $S_Q - 1 = 24$. Отже, перший квартиль становить:

$$Q_1 = x_{Q_1} + h \cdot \frac{\frac{\sum f}{4} - S_{Q_1 - 1}}{f_{Q_1}} = 440 + 20\frac{\frac{200}{4} - 24}{38} = 453{,}7 \ кг.$$

Це означає, що одна чверть корів має живу масу 453,7 кг і менше, а три чверті — більш як 453,7 кг.

Для третього квартиля ¾ частини всіх частот дорівнює 200:4·3 = 150. Цей варіант лежить в інтервалі 480—500. Його нижня межа $x_{Q_3} = 480$, величина $h = 20$, частота $f_{Q_3} = 44$, а сума нагромаджених частот інтервалів, що передують відповідним квартильним $S_{Q_3 - 1} = 125$. Отже, третій квартиль становить:

$$Q_3 = x_{Q_3} + h \cdot \frac{\frac{3}{4}\sum f - S_{Q_3 - 1}}{f_{Q_3}} = 480 + 20\frac{\frac{3}{4} \cdot 200 - 125}{44} = 491{,}4 кг.$$

Це означає, що три чверті корів мають живу масу до 491,4 кг, а одна чверть — більш як 491,4 кг.

Деціли визначають за такою формулою:

$$D = x_D + h\frac{\frac{1}{10}\sum f - S_{D-1}}{f_{D_n}},$$

де x_D — нижня межа відповідного децильного інтервалу; h — величина децильного інтервалу; $S_D - 1$ — сума нагромаджених частот

інтервалів, що передують відповідним децильним; f_D — частоти відповідних децильних інтервалів.

Підставивши дані табл. 15 у формулу, матимемо:

$$D_1 = 420 + 20\frac{\frac{200}{10} - 5}{19} = 435{,}8 \ кг.$$

Отже, десята частина всіх корів має живу масу 435,8 кг і менше, а решта (90 %) — більш як 435,8 кг.

2.5 Показники варіації та способи їх обчислення

Середні величини дають узагальнюючу характеристику сукупностей за якою — небудь ознакою. Проте при незмінному середньому значенні досліджуваної ознаки окремі сукупності істотно різняться за складом і величиною відхилень від середньої. Вивчення розміру відхилень та їх розподілу використовують для оцінки кількісної однорідності сукупності. Чим менше ступінь коливання ознаки, тим однорідніша сукупність.

Вимірювання й аналіз варіації має велике значення для оцінки стійкості досліджуваних явищ, а також впливу різних факторів на коливання ознак.

Варіація — це зміна розміру ознаки у статистичній сукупності. Прикладом варіації може бути зміна розміру урожайності сільськогосподарських культур в окремих господарствах району або області, коливання рівня продуктивності тварин на фермі, коливання розміру заробітної плати у працівників підприємства тощо. Варіація є результатом дії на одиниці сукупності природних, кліматичних, економічних, соціальних та інших факторів, а також індивідуальних особливостей окремих одиниць.

Для характеристики варіації використовують такі показники: розмах варіації, середнє лінійне відхилення, дисперсію, середнє квадратичне відхилення, коефіцієнт варіації тощо. Зміст і порядок визначення показників варіації розглянемо на такому прикладі (табл. 16).

Дані табл. 16 показують, що середній денний виробіток одного орного агрегату в обох орендних колективах однаковий і становить 72:6 = 12 га.

Розмах варіації — це різниця між найбільшим і найменшим значенням варіюючої ознаки:

$$R = x_{max} - x_{min}.$$

Таблиця 16 *Денний виробіток тракторів Т — 150 в агрегаті з плугом ПЛН — 5 – 35 на оранці грунтів, га*

Перший орендний колектив				Другий орендний колектив			
Номер агрегату	Денний виробіток	Відхилення від середньої	Квадрат відхилень	Номер агрегату	Денний виробіток	Відхилення від середньої	Квадрат відхилень
	x	$x-\bar{x}$	$(x-\bar{x})^2$		x	$x-\bar{x}$	$(x-\bar{x})^2$
1	11,8	–0,2	0,04	1	9,9	–2,1	4,41
2	10,9	–1,1	1,21	2	12,7	0,7	0,49
3	13,1	1,1	1,21	3	10,1	–1,9	3,61
4	11,8	–0,2	0,04	4	13,5	1,5	2,25
5	12,0	0	0	5	12,4	0,4	0,16
6	12,4	0,4	0,16	6	13,4	1,4	1,96
Разом	72,0	—	2,66	Разом	72,0	—	12,88

У нашому прикладі розмах варіації деннного виробітку орних агрегатів у першому орендному колективі становитиме 13,1 – 10,9 = = 2,2 га, у другому — 13,5 – 9,9 = 3,6 га.

Розмах варіації дає уявлення лише про межі коливання ознаки, оскільки він ураховує тільки два крайніх значення і не враховує відхилень усіх варіантів.

Для більш точної характеристики варіації ознак окремі їх значення порівнюють з типовим, стійким для сукупності рівнем — величиною середньої. Унаслідок такого порівняння дістають характеристику варіації рядом відхилень від середньої $x-\bar{x}$.

Середнє лінійне відхилення становить середню з абсолютних відхилень усіх варіантів від середнього значення варіюючої ознаки. Його визначають за такими формулами:

$$\bar{d}=\frac{\sum|x-\bar{x}|}{n}-\textit{просте};$$

$$\bar{d}=\frac{\sum|x-\bar{x}|f}{\sum f}-\textit{зважене};$$

У нашому прикладі середнє лінійне відхилення денного виробітку орних агрегатів дорівнюватиме:

$$\bar{d}=\frac{\sum|x-\bar{x}|}{n}=\frac{0{,}2+1{,}1+1{,}1+0{,}2+0+0{,}4}{6}=0{,}5\textit{га};$$

$$\bar{d}=\frac{\sum|x-\bar{x}|}{n}=\frac{2{,}1+0{,}7+1{,}9+1{,}5+0{,}4+1{,}4}{6}=1{,}3\textit{га}.$$

Отже, коливання виробітку у першому орендному колективі менше, порівняно з другим колективом.

Недоліком середнього лінійного відхилення є те, що при його обчисленні не враховуються знаки відхилення. Тому для визначення загального обсягу варіації відхилення від середньої з різними знаками $x-\bar{x}$ підносять до квадрата, підсумовують і одержують як позитивну величину загальний обсяг варіації ознаки в сукупності — су — му квадратів відхилень $\omega=\sum(x-\bar{x})^2$ або з урахуванням частот $\omega=\sum(x-\bar{x})^2 f$.

Обчислення загального обсягу варіації (суми квадратів відхилень) пов'язане із складними розрахунками, особливо коли варіанти або середня величина виражені десятковими дробами. Ці розрахунки можна спростити, якщо використати такі формули обсягу варіації:

$$\omega = \sum (x - \bar{x})^2, \text{ або } \omega = \sum f(x - \bar{x})^2;$$

$$\omega = \sum x^2 - \frac{(\sum x)^2}{n}, \text{ або } \omega = \sum x^2 f - \frac{(\sum xf)^2}{\sum f};$$

$$\omega = \sum x^2 - n\bar{x}^2, \text{ або } \omega = \sum x^2 f - f\bar{x}^2.$$

Залежно від загального обсягу варіації визначають дисперсію та середнє квадратичне відхилення.

Дисперсією називають середній квадрат відхилення всіх значень ознаки від її середньої величини. Її обчислюють за такими формулами:

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n} \text{ — проста;}$$

$$\sigma^2 = \frac{\sum (x - \bar{x})^2 f}{\sum f} \text{ — зважена;}$$

У нашому прикладі (табл. 16) дисперсія становитиме: для першого орендного колективу — 2,66: 6 = 0,44; для другого орендного колективу — 12,88: 6 = 2,15.

Для альтернативних ознак дисперсію визначають за такою формулою:

$$\sigma^2 = \frac{(1-p)^2 p + (0-p)^2 q}{p+q} = q^2 p + p^2 q = pq(p+q) = pq.$$

Отже, дисперсія альтернативної ознаки дорівнює добуткові частки на число, яке доповнює її до одиниці. Оскільки $p + q$ не може бути більше за 1, то σ^2 не може перевищувати 0,25.

Середнє квадратичне відхилення (σ) обчислюють добуванням квадратного кореня з дисперсії:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \text{ — просте;}$$

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2 f}{\sum f}} \text{ — зважене;}$$

$\sigma = \sqrt{pq}$ — для альтернативної ознаки.

У нашому прикладі середнє квадратичне відхилення денного виробітку орних агрегатів дорівнюватиме: для першого орендного колективу — $\sqrt{0,44} = 0,66$ га; для другого орендного колективу — $\sqrt{2,15} = 1,47$ га.

Середнє квадратичне відхилення характеризує середнє коливання ознаки в сукупності, зумовлене індивідуальними особливостями одиниць сукупності. Його виражають у тих самих одиницях вимірювання, що й варіанти досліджуваної ознаки.

Для того щоб порівняти сукупності з різним рівнем середнього значення ознаки та середнього квадратичного відхилення, визначають коефіцієнт варіації (v), який становить відношення середнього квадратичного відхилення до середнього значення ознаки:

$$v = \frac{\sigma}{\bar{x}} \cdot 100.$$

У нашому прикладі коефіцієнт варіації дорівнюватиме: для першого орендного колективу — $\frac{0,66 \cdot 100}{12} = 5,5\%$; для другого орендного колективу — $\frac{1,47 \cdot 100}{12} = 12,2\%$.

Порівняння коефіцієнтів варіації дає змогу зробити висновок, що коливання денного виробітку орних агрегатів у першому орендному колективі значно нижче, ніж у другому.

Якщо варіація в сукупності зумовлена випадковими причинами, то коефіцієнт варіації характеризує відносний вплив випадкових факторів порівняно з основними умовами сукупності, які формують середню величину.

За допомогою коефіцієнта варіації можна порівняти сукупності за коливаннями різних ознак. Припустимо, що середнє квадратичне відхилення продуктивності корів на фермі становить 800 кг при середньому надої молока від однієї корови 3500 кг, а за вмістом жиру в молоці середнє квадратичне відхилення і середня – відповідно 0,56 і 3,7 %. Порівняння коефіцієнтів варіації продуктивності корів $\frac{800}{3500}\cdot 100 = 22{,}9\%$ і жирності молока $\frac{0{,}56}{3{,}7}\cdot 100 = 15{,}1\%$ дає змогу зробити висновок, що продуктивність корів варіює більше, ніж жирність молока.

Послідовність розрахунку показників варіації для інтервального варіаційного ряду розглянемо на такому прикладі (табл. 17).

Середня жива маса однієї корови

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{94400}{200} = 472 \text{ кг.}$$

Таблиця 17 *Розрахунок показників варіації для інтервального варіаційного ряду*

Жива маса корови, кг	Варіант x	Кількість корів F	Розрахункові величини				
			f_x	$x-\bar{x}$	$\lvert x-\bar{x}\rvert f$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
400 – 420	410	5	2050	–62	–310	3844	19220
420 – 440	430	19	8170	–42	–798	164	33516

Продовження таблиці 17

Жива маса корови, кг	Варіант x	Кількість корів F	Розрахункові величини				
			f_x	$x-\bar{x}$	$\lvert x-\bar{x}\rvert f$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
440 – 460	450	38	17100	–22	–836	484	18392
460 –480	470	63	29610	–2	–126	4	252
480 – 500	490	44	21560	18	792	324	14256
500 – 520	510	26	13260	38	988	1444	37544
520 – 540	530	5	2650	58	290	3364	16820
Разом	—	200	94400	—	4140	—	140000

Середнє лінійне відхилення

$$\bar{d} = \frac{\sum\lvert x-\bar{x}\rvert}{\sum f} = \frac{4140}{200} = 20{,}7 \text{ кг.}$$

Дисперсія

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{\sum f} = \frac{140000}{200} = 700.$$

Середнє квадратичне відхилення

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2 f}{\sum f}} = \sqrt{700} = 26{,}5 \ \ кг.$$

Коефіцієнт варіації

$$v = \frac{\sigma}{x} \cdot 100 = \frac{26{,}6 \cdot 100}{472} = 5{,}6\%.$$

2.6 Математичні властивості дисперсії та спрощені способи її обчислення

Дисперсія має деякі математичні властивості, використання яких значно спрощує її розрахунки, зокрема:

1) якщо всі варіанти зменшити або збільшити на яке-небудь постійне число a, то середній квадрат відхилень від цього не зміниться:

$$\sigma^2{}_{x-a} = \sigma^2{}_x.$$

Це означає, що дисперсію можна обчислювати не тільки за варіантами, а й за їх відхиленнями від деякого постійного числа;

2) якщо всі варіанти зменшити або збільшити в a разів, то дисперсія від цього зміниться в a^2 разів, а середнє квадратичне відхилення — в a разів:

$$\sigma^2_{\frac{x}{a}} = \sigma^2_x : a^2.$$

Отже, всі варіанти можна поділити на яке-небудь постійне число (наприклад, на величину інтервалу ряду розподілу), розрахувати середнє квадратичне відхилення, а потім помножити його на це постійне число:

$$\sigma_x = \sigma_{\frac{x}{a}} \cdot a;$$

3) якщо розрахувати середній квадрат відхилень від довільної величини a, яка відрізняється від середньої арифметичної ($\bar{x}$), то він завжди буде більший за середній квадрат відхилень, обчислений від середньої арифметичної:

$$\sigma^2_a > \sigma^2_x.$$

При цьому середній квадрат відхилень значень ознаки x від довільної величини a завжди більший за дисперсію ознаки x на квадрат різниці між середньою арифметичною та довільною величинами:

$$\sigma_a^2 = \sigma_x^2 + (\bar{x} - a)^2, \text{ або } \sigma_a^2 = \frac{\sum (x - a) f}{\sum f} - (\bar{x} - a)^2.$$

З наведених формул видно, що дисперсія від середньої завжди менша за дисперсії, розраховані від будь-яких інших величин, тобто вона має властивість мінімальності.

Якщо довільну величину a прирівняти до нуля, то попередня формула матиме такий вигляд:

$$\sigma^2 = \frac{\sum x^2 f}{\sum f} - \left(\frac{\sum xf}{\sum f}\right)^2 \text{ або } \sigma^2 = \bar{x^2} - \bar{x}^2$$

Отже, дисперсія ознаки дорівнює різниці між середнім квадратом значень ознаки та квадратом їх середнього. Цю формулу широко використовують у статистиці для спрощеного розрахунку дисперсії (табл. 18).

Розрахунок дисперсії за формулою проведемо за формулою:

$$\sigma^2 = \bar{x_2} - \bar{x}^2$$

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{4620}{344} = 13{,}4;$$

$$\overline{x^2} = \frac{\sum x^2 f}{\sum f} = \frac{63040}{344} = 183{,}3;$$

$$\sigma^2 = \overline{x^2} - \bar{x}^2 = 183{,}3 - 13{,}4 = 183{,}3 - 179{,}6 = 3{,}7\ .$$

Таблиця 18 *Спрощений розрахунок дисперсії*

Витрати палива на оранку 1 га, кг х	Площа поля, га, F	Розрахункові дані	
		f_x	$x^2 f$
11	50	550	6050
12	45	540	6480
15	54	810	12150
14	45	630	8820
12	55	660	7920
14	45	630	8820
16	50	800	12800
Разом	344	4620	63040

Використання зазначених вище властивостей значно спрощує обчислення дисперсії, особливо для розрахунків на лічильних машинах. Так, використовуючи другу і третю властивості в рядах розподілу з рівновеликими інтервалами h, дисперсію можна обчислити способом відліку від умовного нуля а (способом моментів). У цьому разі її визначають за формулою

$$\sigma^2 = \frac{\sum \left(\frac{x - a}{h}\right)^2 f}{\sum f} \cdot h^2 - (\bar{x} - a)^2,$$

де a — середина центрального інтервалу.

Середню величину із значень $\frac{x - a}{h}$ називають моментом першого порядку і позначають через m_1, а середній квадрат із значень $\frac{x - a}{h}$ — моментом другого порядку і позначають через m_2. Дисперсія, обчислена за способом моментів, дорівнює квадрату величини

інтервалу, помноженому на різницю моменту другого порядку і квадрата моменту першого порядку:

$$\sigma^2 = h^2\left(m_2 - m_1^2\right)$$

Розрахунок дисперсії способом відліку від умовного нуля розглянемо на такому прикладі (табл. 19).

У нашому прикладі за початок відліку взято усереднений варіант з найбільшою частотою (a = 470). Величина інтервалу h = 20.

Середня жива маса однієї корови становить

$$\bar{x} = \frac{\sum\left(\frac{x-a}{h}\right)f}{\sum f}\cdot h + a = \frac{20}{200}\cdot 20 + 470 = 472 \ \ кг;$$

дисперсія живої маси корів

$$\sigma^2 = \frac{\sum\left(\frac{x-a}{h}\right)^2 f}{\sum f}\cdot h^2 - (\bar{x}-a)^2 = \frac{352}{200}\cdot 20^2 - (472-470)^2 = 700;$$

момент першого порядку

$$m_1 = \frac{\sum\left(\frac{x-a}{h}\right) f}{\sum f} = \frac{20}{200} = 0{,}1;$$

момент другого порядку

$$m_2 = \frac{\sum\left(\frac{x-a}{h}\right)^2 f}{\sum f} = \frac{352}{200} = 1{,}76;$$

дисперсія

$$\sigma^2 = h^2\left(m_2 - m_1^2\right) = 20^2(1{,}76 - 0{,}01) = 700.$$

Таблиця 19 *Розрахунок дисперсії способом відліку від умовного нуля*

Жива маса корови, кг	Кількість корів	Середина інтервалу	Відхилення від умовного нуля	Відхилення в інтервалах	Добуток відхилень в інтервалах на частоти	Квадрати відхилень в інтервалах	Добуток квадратів відхилень в інтервалах на частоти
	f	x	$x - a$	$\frac{x-a}{h}$	$\left(\frac{x-a}{h}\right)f$	$\left(\frac{x-a}{h}\right)^2$	$\left(\frac{x-a}{h}\right)^2 f$
400 – 420	5	410	–60	–3	–15	9	45
420 – 440	19	430	–40	–2	–38	4	76
440 – 460	38	450	–20	–1	–38	1	38
460 – 480	63	470	0	0	0	0	0
480 – 500	44	490	20	1	44	1	44
500 – 520	26	510	40	2	52	4	104
520 – 540	5	530	60	3	55	9	45
Разом	200	—	—	—	20	—	352

Розрахунок дисперсії способом відліку від умовного нуля значно скорочує обсяг обчислювальної роботи тоді, коли значення ознаки виражене у вигляді рівних інтервалів. В інших випадках дисперсію рекомендується розраховувати за такою формулою:

$$\sigma^2 = \overline{x^2} - \overline{x}^2.$$

2.7 Правило складання (розкладання) варіації та його використання у статистичному аналізі

Якщо досліджувану сукупність поділити на групи за факторною ознакою, то для цих груп можна обчислити такі види варіації: загальну, міжгрупову та внутрішньогрупову.

Загальна варіація (ω^3) характеризує вплив усіх умов на варіацію ознаки. Її обчислюють як суму квадратів відхилень індивідуальних значень ознаки x від загальної середньої $\overline{x}_3$.

Міжгрупова варіація ω_M характеризує систематичну варіацію, зумовлену факторною ознакою, за якою проведене групування. Вона дорівнює сумі квадратів відхилень групових середніх $\overline{x}_{гр}$ від загальної середньої $(\overline{x})$. Її визначають за такими формулами:

$$\omega_M = \sum n_{гр}\left(\overline{x}_{гр} - \overline{x}_3\right)^2;$$

$$\omega_M = \frac{\sum\left(\sum x_{гр}\right)^2}{n} - \frac{\left(\sum x\right)^2}{n};$$

$$\omega_M = \frac{\sum\left(\sum x_{гр}\right)^2}{n} - N\overline{x}_3^2;$$

де n – кількість спостережень у кожній групі; N – загальна кількість спостережень.

Внутрішньогрупова варіація ω_B характеризує варіацію, зумовлену факторами, не врахованими при групуванні. Вона дорівнює сумі квадратів відхилень індивідуальних значень ознаки від групових середніх в усіх групах

$$\omega_B = \sum\sum\left(x - \overline{x}_{гр}\right)^2 = \sum x^2 - \frac{\sum\left(\sum x_{гр}\right)^2}{n}.$$

Загальний обсяг варіації ознаки дорівнює сумі міжгрупової та внутрішньогрупової варіації:

$$\omega_З = \omega_М + \omega_В,$$

$$\text{або } \sum(x-\bar{x}_З)^2 = \sum n_{гр}(\bar{x}_{гр}-\bar{x}_З)^2 + \sum\sum(x-x_{гр})^2;$$

$$\text{або } \sum x^2 - \frac{(\sum x^2)}{N} = \frac{\sum(\sum x_{гр})^2}{n} - \frac{(\sum x)^2}{N} + \sum x^2 - \frac{\sum(\sum x_{гр})^2}{n};$$

$$\text{або } \sum x^2 - N\bar{x}_З^2 = \frac{\sum(\sum x_{гр})^2}{n} - N\bar{x}_З^2 + \sum x^2 - \frac{\sum(\sum x_{гр})^2}{n}.$$

Загальна дисперсія ознаки відповідно дорівнюватиме сумі міжгрупової та внутрішньогрупової дисперсій:

$$\sigma_З^2 = \sigma_М^2 + \sigma_В^2.$$

Наведена вище рівність дістала назву правила складання (розкладання) варіації. У статистичному аналізі це правило використовують для оцінки впливу окремих факторів на загальну варіацію ознаки в сукупності. Якщо відомий обсяг загальної та міжгрупової варіацій, то, віднімаючи від загальної варіації міжгрупову, можна визначити обсяг внутрішньогрупової варіації. Порівнюючи обсяги міжгрупової та загальної варіацій, установлюють вплив факторної ознаки, за якою проведене групування, на коливання результативної ознаки.

Порядок визначення загального обсягу варіації та розкладання його на міжгрупову і внутрішньогрупову варіації розглянемо на такому прикладі (табл. 20).

Таблиця 20 *Вплив доз форфорних добрив на урожайність гречки*

Варіант досліду	Урожайність на ділянках, ц/га				Сума $\sum x_{гр}$	Групові середні $\bar{x}_{гр}$
	1	2	3	4		
	x					
№ $_{40}K_{40}$ (фон)	14,9	15,1	16,3	15,7	62,0	15,5
Фон+P_{20}	17,1	15,2	17,0	18,1	67,4	16,9
Фон+P_{40}	16,4	17,4	18,2	18,4	70,4	17,6
Фон+P_{60}	17,8	19,6	18,3	20,3	76,0	19,0
Разом	—	—	—	—	275,8	17,2

Щоб визначити суми квадратів відхилень, показники урожайності гречки піднесемо до квадрата (табл. 21).

Таблиця 21 *Таблиця квадратів*

Варіант досліду	Урожайність на ділянках, ц/га				Сума квадратів	Квадрат суми
	1	2	3	4		
	x^2				$\sum x^2$	$(\sum x_{гр})^2$
№ $_{40}K_{40}$ (фон)	222,01	228,01	265,69	246,49	962,20	3844,00
Фон+P_{20}	292,41	231,04	289,00	327,61	1140,06	4542,76
Фон+P_{40}	268,96	302,76	331,24	338,56	1241,52	4956,66
Фон+P_{60}	316,84	384,16	334,89	412,09	1447,98	5776,00
Разом	—	—	—	—	4791,76	19118,92

Загальний обсяг варіації урожайності гречки

$$\omega_3 = \sum x^2 - \frac{(\sum x)^2}{N} = 4791{,}76 - \frac{275{,}8^2}{16} = 37{,}66;$$

обсяг міжгрупової варіації, зумовленої різними дозами фосфорних добрив,

$$\omega_M = \frac{\sum(\sum x_{гр})^2}{n} - \frac{(\sum x)^2}{N} = \frac{19118{,}92}{4} - \frac{275{,}8^2}{16} = 25{,}63;$$

обсяг внутрішньогрупової варіації

$$\omega_B = \sum x^2 - \frac{\sum(\sum x_{гр})^2}{n} = 4791{,}76 - \frac{19118{,}92}{4} = 12{,}03.$$

Отже, $\omega_3 = \omega_M + \omega_B = 25{,}63 + 12{,}03 = 37{,}66$.

Загальну варіацію урожайності гречки (37,66) поділено на систематичну, зумовлену впливом доз фосфорних добрив (25,63), і випадкову, зумовлену впливом не врахованих у досліді факторів (12,03). Отже, 25,63:37,66·100 = 68,1 % загального коливання урожайності гречки в досліді пояснюється впливом досліджуваного фактора, а 12,03:37,66·100 = 31,9 % — впливом не врахованих випадкових факторів.

2.8 Моменти статистичного розподілу

Варіаційний ряд розподілу характеризується системою узагальнюючих числових показників, які одержали назву моментів статистичного розподілу. Середня арифметична та дисперсія є частковими випадками статистичних моментів.

Моментом розподілу називається середня арифметична величина з ніднесених до заданого ступеня відхилень окремих варіантів ряду розподілу від деякої постійної величини (0, $\bar{x}$, x_0). Загальний математичний вираз моменту розподілу має такий вигляд:

$$M_к = \frac{\sum(x_i - A)^к f_i}{\sum f_i};$$

де $M_к$ — момент к-го порядку; x_i — варіанти ряду розподілу; A — постійна величина, від якої визначаються відхилення (за постійну величину можуть бути взяті нуль, середня арифметична $\bar{x}$ або умовний початок відліку x_0); $к$ — показник ступеня, що визначає порядок моменту; f — частоти ряду.

Для визначення характеристик розподілів найчастіше використовують моменти перших п'яти порядків при значенні K — 0, 1, 2, 3 і 4. Залежно від прийнятого значення постійної величини A розрізняють три види моментів: початкові, центральні й умовні.

Початковими називаються моменти розподілу, при визначенні яких за постійну величину A приймається нуль. У загальному вигляді їх розраховують за формулою:

$$M_к = \frac{\sum x_i^к f_i}{\sum f_i},$$

де $M_к$ — початковий момент $к$-го порядку; x_i^k — варіанти $к$-го ступеня; f_i — частоти.

При $к = 0$ одержують початковий момент нульового порядку, при $к = 1$ — початковий момент першого порядку, при $к = 2$ — початковий момент другого порядку і т.д.

Формули для розрахунку початкових моментів перших п'яти порядків наведені в табл. 22.

Таблиця 22 *Формули та зміст початкових моментів*

Порядок, K	Формула	Зміст
0	$M_0 = \frac{\sum x_i^0 f_i}{\sum f_i}$	1
1	$M_1 = \frac{\sum x_i^1 f_i}{\sum f_i}$	$\bar{x}$ (середня арифметична)
2	$M_2 = \frac{\sum x_i^2 f_i}{\sum f_i}$	$\bar{x^2}$ (середня квадратів варіантів)
3	$M_3 = \frac{\sum x_i^3 f_i}{\sum f_i}$	$\bar{x^3}$ (середня кубів варіантів)
4	$M_4 = \frac{\sum x_i^4 f_i}{\sum f_i}$	$\bar{x^4}$ (середня четвертих ступенів варіантів)

Центральними (M) називаються моменти розподілу, при визначенні яких за постійну величину А приймається середня арифметична. У загальному вигляді розраховують їх за формулою:

$$M_к = \frac{\sum (x_i - \bar{x})^к f_i}{\sum f_i} 3$$

Із наведеної формули видно, що центральні моменти являють собою середні значення різних ступенів відхилень від середньої арифметичної. Формули для розрахунку центральних моментів перших п'яти порядків наведені в табл. 23.

Таблиця 23 *Формули і зміст центральних моментів*

Порядок, K	Формула	Зміст	Взаємозв'язок з початковими моментами
0	$M_0 = \frac{\sum (x_i - \bar{x})^0 f_i}{\sum f_i}$	1	—
1	$M_1 = \frac{\sum (x_i - \bar{x})^1 f_i}{\sum f_i}$	0	$M_1 - M_1$
2	$M_2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}$	σ^2	$M_2 - M_1^2$
3	$M_4 = \frac{\sum (x_i - \bar{x})^3 f_i}{\sum f_i}$	Використовується для визначення показника асиметрії розподілу	$M_3 - 3M_2M_1\ M_1^3$
4	$M_4 = \frac{\sum (x_i - \bar{x})^4 f_i}{\sum f_i}$	Виткористовується для визначення показника гостровершинності розподілу	$M_4 - 4M_3M_1 + 6M_2M_1^2 - 3M_1^4$

Центральні моменти мають деякі математичні властивості, використання яких значно спрощує їх розрахунки, зокрема:

1) якщо всі варіанти ряду розподілу зменшити або збільшити на яке-небудь постійне число C, то величина центрального моменту K-го порядку не зміниться. Це означає, що центральні моменти можна обчислювати не тільки за варіантами, а й за їх відхиленнями від деякого постійного числа;
2) якщо всі варіанти ряду розподілу зменшити або збільшити в одне і теж число раз Z, то центральний момент K-го порядку зменшиться або збільшиться в $Z_к$ разів. Отже, всі варіанти можна поділити на яке небудь постійне число (наприклад, на величину інтервалу ряду розподілу) розрахувати центральний момент для зміненого ряду, а потім помножити його на це постійне число у відповідному ступені.

Умовними (m) називаються моменти розподілу, при визначенні яких за постійну величину А приймається довільно взята величина (x_0), яку називають умовним початком відліку. В загальному вигляді розраховують їх за формулою:

$$m_k = \frac{\sum (x_i - x_0)^k f_i}{\sum f_i}$$

Таким чином, умовні моменти являють собою середні різних ступенів із відхилень варіантів ряду розподілу від умовного початку. Умовні моменти першого, другого, третього і вищих порядків визначають за такими формулами:

$$m_1 = \frac{\sum (x_i - x_0)^1 f_i}{\sum f_i};\ m_2 = \frac{\sum (x_i - x_0)^2 f_i}{\sum f_i};\ m_3 = \frac{\sum (x_i - x_0)^3 f_i}{\sum f_i}$$ і т.д.

Як було показано раніше, умовні моменти першого та другого порядків використовують для спрощених розрахунків середньої арифметичної та дисперсії.

Аналізуючи формули моментів, можна помітити, що початковий момент першого порядку являє собою середню арифметичну і використовується як показник центру розподілу. Центральний момент першого порядку завжди дорівнює нулю, оскільки алгебраїчна сума відхилень варіантів від середньої арифметичної, зважених відповідними частотами, дорівнює нулю. Центральний момент другого порядку дорівнює дисперсії. Центральний момент третього порядку дорівнює нулю в симетричному розподілі і використовується для характеристики асиметрії (скошеності). Центральний момент четвертого порядку використовується для визначення показника ексцесу (гостровершинності).

Нерідко в статистичній практиці більш зручно користуватися не абсолютними значеннями моментів, а їх відношенням до стандартного (середнього квадратичного) відхилення відповідного ступеня. Відношення центрального моменту K-го порядку до середнього квадратичного відхилення в K-му ступені називається нормованим моментом. В загальному вигляді розраховують його за формулою:

$$m'_K = \frac{M_K}{\left(\sqrt{M_2}\right)^K} = \frac{M_K}{\delta^K}$$

Формули для розрахунку нормованих моментів від першого до четвертого порядків мають такий вигляд:

$$m'_1 = \frac{M_1}{\delta^1} = \frac{0}{\delta} = 0;\ m'_2 = \frac{M_2}{\delta^2} = \frac{\delta^2}{\delta^2} = 1;\ m'_3 = \frac{M_3}{\delta^3};\ m'_4 = \frac{M_4}{\delta^4}.$$

Порядок визначення початкових, умовних і центральних моментів розглянемо на прикладі розподілу 50 корів за рівнем молочної продуктивності. Вихідні та розрахункові дані для обчислення початкових моментів наведені в табл. 24.

Таблиця 24 *Вихідні і розрахункові дані для визначення початкових моментів варіаційного ряду розподілу корів за рівнем молочної продуктивності*

Надій молока від корови за рік, ц	Кількість корів	Середина інтервалу	Ступені серединних значень інтервалів			Добуток			
	f_i	x_i	x_i^2	x_i^3	x_i^4	$x_i f_i$	$x_i^2 f_i$	$x_i^3 f_i$	$x_i^4 f_i$
20 — 24	4	22	484	10648	234256	88	1936	42592	937024
24 — 28	13	26	676	17576	456976	338	8788	228488	5940688
28 — 32	19	30	900	27000	810000	570	17100	513000	15390000
32 — 36	11	34	1156	39304	1336336	374	12716	432344	14699696
36 — 40	3	38	1444	54872	2085136	114	4332	164616	6255408
Разом	50	—	—	—	—	1484	44872	1381040	43222816

За даними табл. 24 визначимо початкові моменти перших п'яти порядків:

$$M_0 = \frac{\sum x_1^0 f_i}{\sum f_i} = \frac{\sum f_i}{\sum f_i} = \frac{50}{50} = 1;$$

$$M_1 = \frac{\sum x_1 f_i}{\sum f_i} = \frac{1484}{50} = 29{,}68 \text{ (середня арифметична)}$$

$$M_2 = \frac{\sum x_1^2 f_i}{\sum f_i} = \frac{44872}{50} = 897{,}44$$

$$M_3 = \frac{\sum x_1^3 f_i}{\sum f_i} = \frac{1381040}{50} = 27620{,}8$$

$$M_4 = \frac{\sum x_1^4 f_i}{\sum f_i} = \frac{43222816}{50} = 864456{,}32$$

Вихідні і розрахункові дані для обчислення умовних моментів наведені в табл. 25. За умовний початок відліку прийнято серединне значення інтервалу з продуктивністю корів 30 ц ($x_0 = 30$), який має найбільшу частоту. Величина інтервалу становить 4 ц ($h = 4$).

За даними табл. 25 визначимо значення моментів відносно початку відліку, виражених в частках інтервалу:

$$m_0' = \frac{\sum (x_i - x_0)^0 f_i}{\sum f_i} = \frac{\sum f_i}{\sum f_i} = \frac{50}{50} = 1;$$

$$m_1' = \frac{\sum (x_i - x_0)\ f_i}{\sum f_i} = \frac{-4}{50} = -0{,}08;$$

$$m_2' = \frac{\sum (x_i - x_0)^2 f_i}{\sum f_i} = \frac{52}{50} = 1{,}04;$$

Таблиця 25 *Вихідні і розрахункові дані для визначення умовних моментів варіаційного ряду розподілу корів за рівнем молочної продуктивності*

Середина інтервалу, ц	Кількість корів	Ступені відхилень варіантів від умовного початку відліку в інтервалах				Добуток			
x_i	f_i	$x_i' = \frac{x_i - x_0}{n}$	$(x_i')^2$	$(x_i')^3$	$(x_i')^4$	$x_i' f_i$	$(x_i')^2 f_i$	$(x_i')^3 f_i$	$(x_i')^4 f_i$
22	4	–2	4	–8	16	–8	16	–32	64
26	13	–1	1	–1	1	–13	13	–13	13
30	19	0	0	0	0	0	0	0	0
34	11	1	1	1	1	11	11	11	11
38	3	2	4	8	16	6	12	24	48
Разом	50	—	—	—	—	–4	52	–10	136

$$m_3' = \frac{\sum (x_i - x_0)^3 f_i}{\sum f_i} = \frac{-10}{50} = -0{,}2;$$

$$m_4' = \frac{\sum (x_i - x_0)^4 f_i}{\sum f_i} = \frac{136}{50} = 2{,}72;$$

До визначених моментів внесемо поправку на величину інтервалу ($h = 4$) відповідного ступеня (K) і розрахуємо значення умовних моментів, виражених у вихідній системі одиниць

$$m_k = m_k' h^k$$

$$m_0 = m_k' h_0 = 1;$$

$$m_1 = m_1' h = -0{,}08 \cdot 4 = -0{,}32;$$

$$m_2 = m_2' h^2 = 1{,}04 \cdot 4^2 = 16{,}64;$$

$$m_3 = m_3' h^3 = -0{,}2 \cdot 4^3 = -12{,}8;$$

$$m_4 = m_4' h^4 = 2{,}72 \cdot 4^4 = 696{,}32$$

Визначимо центральні моменти розподілу, користуючись формулами їх взаємозв'язку з початковими моментами:

$$\mu_0 = 1;$$

$$\mu_1 = 0;$$

$$\mu_2 = M_2 - M_1^2 = 897{,}44 - 29{,}682 = 16{,}54;$$

$$\mu_3 = M_3 - 3M_2M_1 + 2M_1^3 = 27620{,}8 - 3 \cdot 897{,}44 \cdot 29{,}68 + 2 \cdot 29{,}683 = 3{,}108;$$

$$\mu_4 = M_4 - 4M_3M_1 + 6M_2M_1^2 - 3M_1^4 = 864456{,}32 - 4 \cdot 27620{,}8 \cdot 29{,}68 + 6 \cdot 897{,}44 \cdot 29{,}682 - 3 \cdot 29{,}684 = 690{,}128.$$

Для перевірки правильності розрахунків визначимо центральні моменти через умовні, користуючись формулами взаємозв'язку між моментами:

$$\mu_0 = 1;$$
$$\mu_1 = 0;$$
$$\mu_2 = m_2 - m_1^2 = 16{,}64 - (-0{,}32) = 16{,}54;$$
$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3 = -12{,}8 - 3 \cdot 16{,}64(-0{,}32) + 2(-0{,}32)3 = 3{,}108;$$
$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4 = 696{,}32 - 4(-12{,}8) \cdot (-0{,}32) + 6 \cdot 16{,}64(-0{,}32)2 - 3(-0{,}32)4 = 690{,}128.$$

Порівняння центральних моментів, визначених через їх взаємозв'язок з початковими і умовними моментами, показує, що обидва ці способи дали однакові результати.

2.9 Характеристика скошеності і гостровершинності розподілу

Статистичний ряд розподілу графічно можна зобразити за допомогою відповідної кривої, яку називають кривою розподілу. Форми таких різноділів залежать від сукупної дії факторів, що впливають на формування значень досліджуваних ознак.

Нормальним називається симетричний розподіл, в якому максимуми значень досліджуваної ознаки концентруються навколо середньої величини. Його особливістю є те, що чим більше відхиляється значення окремого варіанта від середньої, тим рідше він повторюється і навпаки, чим ближче варіанти до середнього значення, тим частіше вони повторюються в даній сукупності. Крива нормального розподілу симетрична відносно осі ординат і асимптотично наближається до осі абсцис.

Як правило, емпіричний (фактичний) розподіл у тій чи іншій мірі відрізняється від нормального. Для оцінки відхилення емпіричного розподілу від нормального розраховують коефіцієнти скошеності (асиметрії) та гостровершинності (ексцесу).

Коефіцієнт скошеності (A_s) обчислюють як відношення центрального моменту третього порядку до куба середнього квадратичного відхилення:

$$A_s = \frac{\mu_3}{\sigma^3}$$

Із цієї формули випливає, що коефіцієнт скошеності являє собою нормований момент третього порядку (m_3).

У наведеному в попередньому параграфі варіаційному ряду розподілу корів за рівнем молочної продуктивності центральний момент третього порядку дорівнює 3,108, а середнє квадратичне відхилення — 4,07. Підставивши значення μ_3 (3,108) і σ (4,07) у формулу коефіцієнта скошеності, маємо:

$$A_s = \frac{\mu_3}{\sigma^3} = \frac{3{,}108}{4{,}07^3} = \frac{3{,}108}{67{,}419} = 0{,}046.$$

Якщо коефіцієнт скошеності дорівнює нулю, то розподіл симетричний, якщо не дорівнює нулю — асиметричний. У випадках, коли $A_s > 0$, розподіл має правосторонню асиметрію, коли $A_s < 0$ — лівосторонню асиметрію.

При симетричному розподілі середня арифметична $(\bar{x})$, мода (M_0) і медіана (Me) рівні між собою. Для правосторонньої (додатньої) скошеності, коли права вітка кривої розподілу довша лівої, існує таке співвідношення між цими статистичними величинами: $M_0 < Me < \bar{x}$. При лівосторонній (від'ємній) скошеності, коли ліва вітка кривої розподілу довша правої, спостерігається обернене співвідношення: $M_0 > M_e > \bar{x}$.

У статистичній практиці прийнято вважати, що при значенні коефіцієнта $A_s < \pm 0{,}25$ асиметрія є незначною, при значенні $A_s > \pm 0{,}5$ — емпіричний розподіл відрізняється від нормального значним зміщенням.

У нашому прикладі коефіцієнт скошеності дорівнює 0,046, що свідчить про правосторонню асиметрію розподілу корів за рівнем молочної продуктивності. Іншими словами, переважаюча кількість корів має молочну продуктивність вищу середньої арифметичної. Абсолютне значення $A_s = 0{,}046 < 0{,}25$ вказує на наявність незначної асиметрії, тобто розподіл досліджуваної групи корів за рівнем молочної продуктивності наближається до нормального.

Для характеристики гостровершинності розподілу використовують нормований момент четвертого порядку, який являє собою відношення центрального моменту четвертого порядку (M_4) до середнього квадратичного відхилення в четвертому ступені (σ_4). При нормальному розподілі нормований момент четвертого порядку дорівнює $3(M_4 / \sigma_4 = 3)$.

Якщо прийняти нормальний розподіл за базу порівняння, то коефіцієнт гостровершинності (ексцесу) можна розрахувати за формулою:

$$E_x = \frac{M_4}{\sigma_4} - 3$$

При нормальному розподілі $E_x = 0$, при гостровершинному (вершина фактичного розподілу виступає над вершиною нормального розподілу) $E_x > 0$, при плосковершинному (вершина фактичного розподілу знаходиться нижче вершини нормального розподілу) $E_x < 0$.

У тих випадках, коли величина коефіцієнта ексцесу не перебільшує ±0,4, крива фактичного розподілу вважається слабоексцесивною. Максимальне значення від'ємного ексцесу становить –2. У цьому випадку вершина кривої фактичного розподілу опускається до осі абсцис і крива розподілу ділиться на дві самостійні одновершинні криві.

У наведеному в попередньому параграфі варіаційному ряду розподілу корів за рівнем молочної продуктивності центральний момент четвертого порядку дорівнює 690,128, а середнє квадратичне відхилення — 4,07. Підставивши значення M_4 (690,128) і σ (4,07) у формулу коефіцієнта гостровершинності, маємо:

$$E_x = \frac{M_4}{\sigma_4} - 3 = \frac{690{,}128}{4{,}07^4} - 3 = -0{,}485.$$

Отже, у нашому випадку, крива розподілу корів за рівнем молочної продуктивності характеризується плосковершинністю з помітновираженою ексцесивністю.

Контрольні питання

1. *Поняття про статистичні ряди розподілу.*
2. *Види статистичних рядів розподілу.*
3. *Із яких елементів складаються статистичні ряди?*
4. *Порядок побудови інтервального ряду розподілу.*
5. *Графічне зображення рядів розподілу.*
6. *Форми статистичних розподілів.*
7. *Нормальний розподіл.*
8. *Асиметрія і ексцес у розподілі.*
9. *Поняття про центральну тенденцію у розподілі.*
10. *Види та форми середніх.*
11. *Основні математичні властивості середньої арифметичної.*
12. *Мода і медіана.*
13. *Суть способів моментів.*
14. *Основні показники варіації.*
15. *Математичні властивості дисперсії.*
16. *Правило складання (розкладання) варіації.*
17. *методів математичної статистики.*

Розділ 3
ВИБІРКОВИЙ МЕТОД

3.1 Теоретичні основи вибіркового методу

За повнотою охоплення одиниць досліджуваної сукупності статистичне спостереження поділяють на суцільне і несуцільне. Суцільне спостереження забезпечує найбільш повну інформацію про загальну кількість одиниць сукупності і дає вірогідні узагальнюючі статистичні характеристики явищ, що аналізуються. Проте здебільшого суцільне спостереження недоцільне або його не можна провести. Так, для визначення втрат при збиранні урожаю суцільне спостереження потребує значних затрат часу та коштів, а при перевірці якості продукції (наприклад, жирності молока, схожості зерна) його не можна провести, оскільки таке обстеження пов'язане із псуванням продукції. У такому разі здійснюють **несуцільне статистичне спостереження**. Найбільш поширеним видом несуцільного спостереження є вибіркове (репрезентативне) спостереження.

Вибірковим називають таке **спостереження**, при якому закономірності і характеристики, властиві якій-небудь генеральній сукупності, визначають дослідженням деякої її частини. Сукупність математичних засобів і обгрунтувань, які використовують при застосуванні вибіркового спостереження, дістала назву **вибіркового методу**.

Теоретичною основою вибіркового методу є закон великих чисел і теорія ймовірності, згідно з якими відмінності між аналогічними характеристиками генеральної та вибіркової сукупностей можна зменшити із збільшенням обсягу вибірки. Вірогідна оцінка всієї досліджуваної сукупності за результатами вибіркового спостереження можлива лише за відповідних умов:

1) кількість відібраних одиниць для спостереження повинна бути досить великою;
2) відбір одиниць для вибіркового спостереження має бути таким, щоб кожна одиниця сукупності мала однакову можливість потрапити у вибірку.

Правильно організоване вибіркове спостереження має кілька істотних переваг перед суцільним. Воно дає змогу дістати потрібні дані з меншими затратами часу та коштів, тобто є економічнішим порівняно із суцільним спостереженням, а також дає змогу швидше підбивати підсумки і робити відповідні висновки, тобто воно є оперативнішим. Вибіркове спостереження застосовують і тоді, коли суцільне спостереження неможливе, наприклад, під час контролю за якістю продукції, при дослідженні споживання населенням, спостереженні сукупностей великих обсягів тощо. Іноді, коли помилки реєстрації значні, вибіркове спостереження забезпечує точніші результати.

У статистичній практиці вибіркове спостереження застосовують при обстеженні домогосподарств населення, для обліку цін на ринках, для визначення втрат при збиранні урожаю, контролю якості продукції тощо. Останнім часом вибіркове спостереження широко застосовують при різних опитуваннях громадської думки з політичних, економічних, соціальних і комерційних питань, у науковій роботі при статистичній обробці результатів досліджень.

Вибіркове спостереження також застосовують у поєднанні із суцільним для поглиблення дослідження (наприклад, при переписах населення) або для перевірки результатів суцільного спостереження. Так, під час перепису (обліку) худоби на основі вибірки визначають процент недообліку при суцільному переписі худоби у населення.

Розрізняють генеральну та вибіркову сукупності. **Генеральна сукупність** — це загальна сукупність одиниць, з якої проводять відбір частини одиниць. **Вибіркова сукупність** — це частина генеральної сукупності, яку вибірково обстежуватимуть. Обсяг генеральної сукупності позначають через N, а вибіркової — через n.

За допомогою вибіркового спостереження вивчають середній розмір досліджуваної ознаки (наприклад, середній розміру втрат продукції при збиранні урожаю) або частку досліджуваної ознаки

(наприклад, частку посівів, уражених шкідниками та хворобами сільськогосподарських культур).

Середню величину та дисперсію ознаки у генеральній сукупності називають **генеральною середньою** і **генеральною дисперсією**. Генеральну середню позначають через $\bar{x}$, а генеральну дисперсію — через σ_0^2.

Середню величину та дисперсію ознаки у вибірковій сукупності називають **вибірковою середньою** і **вибірковою дисперсією**. Вибіркову середню позначають через $\tilde{x}$, а вибіркову дисперсію — через σ_0^2.

Частку одиниць, які мають певні ознаки у генеральній сукупності, називають **генеральною часткою**, або просто **часткою**, а частку одиниць, які мають певні ознаки у вибірковій сукупності, називають **частістю**. Генеральну частку позначають через p, а частість — через w.

Обчислимо показники генеральної та вибіркової сукупностей на такому прикладі. Припустімо, що орендний колектив доглядає 100 корів. Для визначення середньої жирності молока, а також частки корів із жирністю 3,7 % і більше спочатку візьмемо вибіркові проби від 30 корів, а потім – від 100 корів. Жирність молока буде такою, (%):

- від 30 корів: 3,6; 4,0; 3,9; 3,5; 3,3; 3,8; 3,7; 3,8; 3,5; 3,4; 3,8; 3,7; 3,8; 3,5; 3,6; 3,4; 4,0; 3,9; 3,5; 3,3; 3,5; 3,3; 3,9; 3,4; 4,0; 3,6; 3,5; 3,8; 3,7; 3,8;
- від 100 корів: 3,4; 3,8; 3,3; 3,6; 3,5; 3,7; 3,8; 3,7; 3,5; 3,4; 3,5; 3,7; 4,0; 3,9; 3,5; 3,8; 3,7; 3,3; 3,6; 3,7; 3,4; 3,8; 3,5; 3,3; 3,8; 3,6; 3,7; 3,5; 3,9; 3,5; 3,7; 3,5; 4,0; 3,9; 3,5; 3,8; 3,6; 3,7; 3,3; 3,7; 3,5; 3,8; 3,4; 3,7; 3,7; 3,5; 3,6; 3,3; 3,4; 3,8; 3,8; 3,5; 3,4; 3,3; 3,8; 3,6; 3,7; 3,5; 3,5; 3,9; 3,5; 3,8; 3,4; 3,3; 3,8; 3,6; 3,5; 3,7; 3,9; 3,5; 3,7; 3,5; 3,9; 4,0; 3,8; 3,5; 3,7; 3,3; 3,6; 3,7; 3,4; 3,7; 3,5; 3,8; 3,7; 3,5; 3,6; 3,8; 3,3; 3,4; 3,2; 3,6; 4,0; 4,1; 3,6; 3,3; 3,9; 3,6; 3,7; 3,0.

За наведеними даними визначимо показники вибіркової сукупності:

- середня жирність молока $\tilde{x} = \frac{\sum x}{n} = \frac{109.5}{30} = 3{,}65$;
- дисперсія жирності молока $\sigma_0^2 = \frac{\sum (x - \tilde{x})^2}{n} = \frac{1{,}40}{38} = 0{,}047$;
- частка корів з жирністю молока 3,7 % і більше $\omega = \frac{m}{n} = \frac{15}{30} = 0{,}50$.

Показники генеральної сукупності становитимуть:

- середня жирність молока $\bar{x} = \frac{\sum x}{N} = \frac{360}{100} = 3{,}60$;
- дисперсія жирності молока $\sigma^2 = \frac{\sum (x - \bar{x})^2}{N} = \frac{4{,}3}{100} = 0{,}043$;
- частка корів з жирністю молока 3,7 % і більше $p = \frac{M}{N} = \frac{46}{100} = 0{,}46$.

Отже, вибіркова середня дорівнює 3,65 %, вибіркова дисперсія — 0,047, а частість — 0,50; відповідно генеральна середня — 3,60 %, генеральна дисперсія — 0,043, а частка — 0,46. Як видно, ті самі показники вибіркової та генеральної сукупностей не збігаються. В абсолютному виразі різниця між середніми дорівнює 0,05 %, а між частістю і часткою — 0,04.

Різниця між показниками вибіркової та генеральної сукупностей $\left(\bar{x} - \tilde{x}; \sigma_0^2 - \sigma^2; \omega - p\right)$ становить помилку репрезентативності.

Помилки репрезентативності виникають тому, що вибіркова сукупність неточно відображує генеральну сукупність. Ці помилки характерні для всіх вибіркових спостережень, оскільки як би старанно і правильно не проводився відбір одиниць, середні та відносні показники вибіркової сукупності завжди будуть якоюсь мірою відрізнятися від відповідних показників генеральної сукупності. У зв'язку з цим основним завданням вибіркового методу є одержання таких вибіркових характеристик, які б найбільш точно відображували характеристики генеральної сукупності.

Наближене значення шуканої величини генеральної сукупності, встановлене на основі вибіркового спостереження, називають **статистичною оцінкою параметра розподілу**. Основними вимогами, які ставляться до статистичних оцінок, є незміщеність, надійність, ефективність і достатність.

Незміщеною називають таку **статистичну оцінку**, математичне сподівання (можливе значення генеральної характеристики) якої при будь-якому обсязі вибірки дорівнює значенню параметра генеральної сукупності. Оцінка буде незміщеною у тому разі, якщо значення вибіркової характеристики збігається із значенням генеральної характеристики. Якщо ж вибіркова характеристика більша або менша за генеральну, то оцінка буде **зміщеною**.

Надійною називають таку **статистичну оцінку**, яка грунтується на законі великих чисел, тобто із збільшенням кількості спостережень вона наближається до свого математичного сподівання. Надійність оцінки означає, що, чим більша чисельність вибірки, тим вища ймовірність, що помилка оцінки не перевищить скільки завгодно малої величини.

Ефективною називають таку **незміщену оцінку**, яка має найменшу дисперсію серед усіх можливих незміщених оцінок параметра, обчислених за вибірками того самого обсягу. Вона характеризує якісний бік вибіркових характеристи: з двох оцінок ефективною буде та, яка матиме найменшу помилку.

Оцінку називають **достатньою**, якщо вона забезпечує повноту використання всієї інформації про невідому характеристику генеральної сукупності, яка міститься у вибірці.

У теоретичному курсі математичної статистики доводиться, що вибіркова середня арифметична є незміщеною, надійною, ефективною та достатньою оцінкою генеральної середньої, а вибіркова дисперсія є найкращою оцінкою генеральної дисперсії. Математична надійність вибіркової дисперсії не дорівнює дисперсії генеральної сукупності. Через це вибіркова дисперсія є зміщеною оцінкою генеральної дисперсії. При цьому помилка зміщення дорівнює $\frac{n}{n-1}$

Чим менша вибірка, тим більше помилка.

Для того щоб усунути помилку зміщення, вибіркову дисперсію коригують на поправочний коефіцієнт (поправка Бесселя):

$$S^2 = \sigma_0^2 \frac{n}{n-1} = \frac{\sum(x-\tilde{x})^2}{n} \bullet \frac{n}{n-1} = \frac{\sum(x-\tilde{x})^2}{n-1},$$

де S^2 — виправлена вибіркова дисперсія; $\sigma_0{}^2$ — фактична вибіркова дисперсія; $\tilde{x}$ — вибіркова середня арифметична; n — вибіркова сукупність.

Математична надійність виправленої дисперсії при будь-якому обсязі вибірки дорівнює генеральній дисперсії.

Невідомий параметр генеральної сукупності можна оцінити одним числом (точкою) або деяким інтервалом, в якому з певною ймовірністю може міститися шуканий параметр. У зв'язку із цим розрізняють два способи оцінки параметрів генеральної сукупності: точкову та інтервальну оцінки.

Суть **точкової оцінки** полягає в тому, що за певний параметр генеральної сукупності беруть знайдене за вибіркою його числове значення, тобто шуканий параметр оцінюють одним числом. Так, вибіркова середня є незміщеною та найбільш ефективною точковою оцінкою генеральної середньої, а вибіркова виправлена дисперсія – незміщеною точковою оцінкою генеральної дисперсії.

Оскільки вибіркова оцінка є випадковою величиною і має імовірний характер, то числову характеристику слід доповнити величиною середньої помилки. Середня помилка показує можливі відхилення характеристик вибіркової сукупності від характеристик генеральної сукупності. Розмір помилки оцінки залежить від величини дисперсії. Чим менша дисперсія, тим менша помилка оцінки і тим надійніші статистичні висновки.

Теоретично з кожної генеральної сукупності можна сформувати всі можливі вибірки, тобто можливі поєднання одиниць генеральної сукупності при встановленій чисельності вибірок. Якщо генеральна сукупність містить N елементів, а для обстеження

потрібно вибрати з них частину n, то число можливих вибірок визначається за формулою:

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

Усі вони мають однакову ймовірність $\frac{1}{C_N^n}$, але кожна з них несе в собі певну похибку, що відбиває факт випадковості вибору. Оскільки помилки вибіркового спостереження носять випадковий характер, то вони можуть мати різні значення. У наведеному вище прикладі помилка середньої жирності молока у вибірковій сукупності дорівнює 0,05 %. Проте якби у вибірку потрапило більше корів з низькою або високою жирністю, то вибіркова середня могла б бути нижчою або вищою за генеральну середню більш як на 0,05 %.

Для узагальнюючої характеристики помилки вибірки визначають середню помилку. **Середня помилка вибірки** — це середнє квадратичне відхилення вибіркових середніх від середньої генеральної сукупності. У теорії ймовірності доведено, що квадрат середньої помилки, тобто дисперсія вибіркових середніх, прямо пропорційний дисперсії ознаки в генеральній сукупності і обернено пропорційний чисельності вибірки:

$$m^2 = \frac{\sigma^2}{n}.$$

Звідси формула для визначення середньої помилки вибірки матиме такий вигляд:

$$m = \sqrt{\frac{\sigma^2}{n}}, \text{ або } m = \frac{\sigma}{\sqrt{n}}.$$

Наведена формула свідчить, що, чим більша чисельність вибірки, тим менший розмір можливої помилки і, навпаки, чим більший рівень варіації досліджуваної ознаки у генеральній сукупності, тим більший розмір можливої помилки.

Отже, якщо середня обчислена за вибірковими даними, то її відхилення від генеральної середньої дорівнюватиме $\pm m$. У нашому прикладі вибіркова сукупність корів n = 30, а дисперсія жирності молока σ^2 = 0,043. Підставивши ці дані у наведену вище формулу, обчислимо середню помилку при визначенні жирності молока у вибірковій сукупності

$$m = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{0{,}043}{30}} = 0{,}04\%.$$

Якщо вибіркове спостереження застосовують для визначення частки досліджуваної ознаки, то середню помилку вибірки обчислюють за формулою

$$m = \sqrt{\frac{pq}{n}},$$

де pq — дисперсія альтернативної ознаки.

У нашому прикладі (p = 0,46) середня помилка при визначенні вибіркової частості корів із жирністю молока 3,7 % і більше

$$m = \sqrt{\frac{0{,}46 \cdot 0{,}54}{30}} = 0{,}09.$$

Середню помилку вибірки використовують для визначення можливих відхилень показників вибіркової сукупності від відповідних показників генеральної сукупності. Однак можна

стверджувати, що генеральні показники не виходять за межі середньої помилки вибірки не з абсолютною вірогідністю, а лише з певним ступенем ймовірності. У зв'язку з цим поряд із середньою помилкою (точковою оцінкою) визначають граничну помилку вибірки, тобто дається інтервальна оцінка параметрів генеральної сукупності.

Інтервальною називають **оцінку**, яка визначається двома числами — кінцями інтервалу, в якому із встановленою імовірністю міститься шуканий параметр. Центром такого інтервалу звичайно є вибіркова оцінка точки, а визначення меж інтервалу пов'язане із середньою помилкою оцінки та надійною ймовірністю. Отже, інтервальна оцінка є доповненням і розширенням точкової оцінки відповідного параметра. Якщо при точковій оцінці надійність висновку пов'язана тільки з середньою помилкою, то при інтервальній оцінці — з граничною помилкою.

Конкретна помилка ε (різниця між показниками вибіркової та генеральної сукупностей) кожної вибірки може мати різні значення. Проте якщо вибіркова сукупність n досить велика і розподілена нормально, то її відношення до середньої помилки здебільшого не перевищує ± 3. Ця залежність між граничною та середньою помилками у статистиці дістала назву **правила трьох сигм**.

Відношення помилки вибірки до середньої помилки називають **нормованим відхиленням** і позначають через t:

$$t = \frac{\varepsilon}{m}.$$

Підставивши у формулу нормованого відхилення значення ε, матимемо:

$$t = \frac{\tilde{x} - \bar{x}}{m}.$$

Нормоване відхилення використовують як коефіцієнт надійності, який показує, з якою імовірністю можна твердити про значення граничної помилки.

Гранична помилка — це t разів узята середня помилка:

$$\Delta = tm,$$

де m — середня помилка вибірки; t — нормоване відхилення (коефіцієнт надійності).

Нормоване відхилення залежить від ймовірності, з якою можна гарантувати певні розміри граничної помилки. Ймовірність — це міра об'єктивної можливості здійснення певних подій. Кількісно ймовірність виражають відношенням кількості сприятливих наслідків до кількості можливих наслідків. Якщо кількість сприятливих наслідків дорівнює нулю, то ймовірність (p) також дорівнює нулю. Якщо кількість сприятливих наслідків дорівнює всім можливим наслідкам, то ймовірність дорівнює одиниці. Отже, ймовірність може мати значення від 0 до 1.

Теоретичне обгрунтування формули граничної помилки вибірки наведено у відомій теоремі П. Л. Чебишева з доповненнями О. М. Ляпунова. Ця теорема формулюється так: при досить великій кількості незалежних спостережень з імовірністю, як завгодно близькою до одиниці, можна твердити, що вибіркова середня буде як завгодно мало відрізнятися від генеральної середньої.

Стосовно вибіркового спостереження, при якому вивчається середнє значення ознаки, цю теорему можна записати так:

$$p\left[(\tilde{x} - \bar{x}) \le t\sqrt{\frac{\sigma^2}{n}}\right] = \phi(t),$$

де p — ймовірність; Φ — функція.

Для вибіркового спостереження, при якому вивчається частка досліджуваної ознаки, ця формула матиме такий вигляд:

$$p\left[(\omega - p) \le t\sqrt{\frac{pq}{n}}\right] = \Phi(t).$$

Наведена формула виражає таку залежність: з імовірністю, як завгодно близькою до одиниці, можна твердити, що при досить великій кількості незалежних спостережень частість буде як завгодно мало відрізнятися від частки.

Теоретичні величини t і p, обчислені на основі стандартної кривої нормального розподілу, наведено в спеціальних таблицях «Значення інтеграла імовірностей при різних значеннях t» (дод. 1). За цією таблицею, знаючи нормоване відхилення t, можна визначити ймовірність або, навпаки, на основі певної ймовірності визначити величину t. У таблиці на перетині рядків і граф розміщуються значення ймовірності, які відповідають певному значенню t. Оскільки в таблиці всі значення ймовірності збільшені в 10 000 разів, то до табличного значення р потрібно дописувати «нуль цілих». Так при $t = 2$ ймовірність р дорівнює 0,9545. Це означає, що з імовірністю 0,9545 можна гарантувати, що розмір граничної помилки не перевищить двократної середньої помилки, тобто висновок про розмір граничної помилки вибірки буде правильним у 9545 випадках із 10 000. В останніх 455 випадках фактична помилка може бути більшою за встановлені розміри.

За даними таблиці (дод. 1) визначають і величину t при відповідному, достатньому для кожного конкретного випадку, рівні ймовірності.

Результати вибіркового спостереження можна оцінити з різним рівнем ймовірності. Рівень ймовірності, який беруть при обчисленні помилки вибіркового спостереження, називають **гарантованим**. У статистичній практиці достатнім рівнем ймовірності

вважається $p = 0{,}95$. Це означає, що тільки у 5 випадках із 100 помилка може перевищувати встановлені розміри.

При більш суворому підході до оцінки результатів і вищій вимогливості до надійності висновків рівень ймовірності підвищується до 0,99, а в особливо відповідальних випадках (наприклад, при оцінці ефективності нових пестицидів) – до 0,999.

Можливий і інший підхід до інтерпретування результатів вибіркового спостереження, зокрема при оцінці ймовірності того, що знайдені під час вибіркового спостереження розбіжності є випадковими, неістотними. У такому разі визначають величину $1-p$, яку називають **рівнем істотності**.

Рівень істотності показує ймовірність, з якою гіпотеза, що перевіряється, може дати помилковий результат. Уважається достатнім рівень істотності 0,05 (або 5 % рівень), а в більш відповідальних спостереженнях — 0,01 і навіть 0,001 (або відповідно 1 % і 0,1 % рівень).

У наведеному вище прикладі граничні помилки при $t = 2$ дорівнюють:

- середньої жирності молока:

$$\Delta = tm = 2 \cdot 0{,}04 = 0{,}08\ \%;$$

- частості корів з жирністю молока 3,7 % і більше:

$$\Delta = tm = 2 \cdot 0{,}09 = 0{,}18.$$

На основі граничної помилки будується надійний інтервал. Інтервальна оцінка параметра

$$\Theta = \widetilde{\Theta} \pm \Delta,$$

а надійний інтервал має вигляд

$$I_p = \left(\tilde{\Theta} - \Delta; \tilde{\Theta} + \Delta\right),$$

де Θ — значення генеральної характеристики; $\tilde{\Theta}$ — значення вибіркової характеристики; Δ — гранична помилка вибірки.

У нашому прикладі надійний інтервал становитиме:

- середньої жирності молока:

$$I_p = (3{,}65 - 0{,}08;\ 3{,}65 + 0{,}08) = (3{,}57;\ 3{,}73);$$

- частості корів з жирністю молока 3,7 % і більше:

$$I_p = (0{,}50 - 0{,}09;\ 0{,}50 + 0{,}09) = (0{,}41;\ 0{,}59).$$

Це означає, що середня жирність молока у генеральній сукупності буде в межах від 3,57 до 3,73 %, а частка корів з жирністю молока 3,7 % і більше — в межах від 41 до 59 %. Знайдені результати гарантуються в 95 випадках із 100.

При вибірковому спостереженні не можна обчислити генеральну дисперсію σ^2. У зв'язку із цим її замінюють вибірковою дисперсією σ^2_0, враховуючи, що співвідношення між ними, як було показано вище, визначають за такою формулою:

$$\sigma_0^2 = \frac{n}{n-1}\sigma^2, \text{ звідси } \sigma^2 = \sigma_0^2 \frac{n}{n-1}.$$

Якщо вибіркова сукупність має велику кількість спостережень, то величина $\frac{n}{n-1}$ наближається до 1 і цю поправку можна не враховувати. В цьому разі вважається, що обчислена вибіркова дисперсія достатньо точно характеризує варіацію ознаки в генеральній сукупності.

У нашому прикладі середня помилка, обчислена з урахуванням вибіркової дисперсії, дорівнюватиме:

- середньої жирності молока

$$m = \sqrt{\frac{\sigma_0^2}{n}} = \sqrt{\frac{0{,}047}{30}} = 0{,}04\%;$$

- частості корів з жирністю молока 3,7 % і більше

$$m = \sqrt{\frac{0{,}50 \cdot 0{,}50}{30}} = 0{,}09.$$

Використання поправочного коефіцієнта $\frac{n}{n-1}$ не змінить знайдених результатів.

Ці розрахунки показують, що розмір середньої помилки (граничної помилки та надійного інтервалу) при використанні вибіркової дисперсії залишився таким самим, як і при використанні генеральної дисперсії.

3.2 Закони розподілу вибіркових характеристик

Дія закону великих чисел виявляється у вигляді певних закономірності у розподілі варіантів за значеннями варіюючої ознаки. Закон розподілу — це співвідношення між можливими значеннями ознаки величин і відповідними ймовірностями. У статистичній практиці найчастіше застосовують такі закони розподілу:

- нормальний розподіл;
- критерій *t*-Стьюдента;
- критерій χ^2 Пірсона;
- *F*-розподіл Фішера.

Нормальний розподіл має більшість ознак соціально-економічних, сільськогосподарських і біологічних об'єктів з безперервним характером варіювання. У зв'язку з цим його беруть за норму будь-якого масового випадкового вивчення ознак.

Нормальний розподіл — це симетричний розподіл, у якому максимуми значень випадкової величини концентруються навколо середньої величини. Його особливістю є те, що, чим більше відхиляється значення окремого варіанта від середньої, тим рідше він повторюється і тим менша ймовірність появи його в генеральній сукупності. І навпаки, чим ближче варіанти до середнього значення, тим частіше вони повторюються і тим більша ймовірність їх появи. Отже, частота відхилень від середнього значення ознаки є функцією їх величини. Нормальний розподіл буває тоді, коли на величину ознаки явища діють випадкові залежні або слабо залежні фактори, кожний з яких відіграє значну роль у загальному підсумку.

Крива нормального розподілу описується рівнянням Гаусса-Лапласа:

$$y = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{t^2}{2}},$$

де y — ордината кривої (теоретична частота); σ_o — середнє квадратичне відхилення генеральної сукупності, яке при практичних розрахунках замінюють вибірковим; t — нормоване відхилення; π — постійна величина, що характеризує відношення довжини кола до його діаметра = 3,14; e – основа натуральних логарифмів (число Ейлера) = 2,718.

Якщо середнє квадратичне відхилення $\sigma_o = 1$, то рівняння матиме такий вигляд:

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

де $f(t)$ — функція щільності нормального розподілу.

Це рівняння називають **стандартним рівнянням нормальної кривої**, а нормальну криву — **нормованою кривою**. Графік щільності нормального розподілу називають нормальною кривою, або кривою Гаусса.

Крива нормального розподілу симетрична відносно осі ординат і асимптотично наближається до осі абсцис. Площа, обмежена кривою та віссю абсцис, дорівнює одиниці. Значення щільності ймовірності $f(x)$ залежить тільки від величини нормованого відхилення t, оскільки π і e – постійні величини. При $t = 0$ $e^{-\frac{t^2}{2}} = 1$ і щільність ймовірності максимальна — $f(o)$ = 0,3989. Із збільшенням t щільність ймовірності зменшується. Щоб установити ймовірність, в інтервалі від t_1 до t_2, потрібно знайти відношення частини площі кривої, що міститься між ординатами t_1 і t_2, до всієї площі кривої. Площу нормальної кривої між ординатами t_1 і t_2 визначають інтегруванням наведеної вище функції. Значення інтеграла імовірностей при відомому t і значення t при вставленому рівні ймовірності можна визначити за спеціальними таблицями (дод. 1). Теоретичні значення t і p, визначені за стандартним рівнянням нормальної кривої, використовують при вибірковому спостереженні як нормативи (критерії), за допомогою яких оцінюють вибіркові харак-

теристики. У зв'язку із цим нормоване відхилення кривої нормального розподілу дістало назву *t*-критерію розподілу нормальної кривої.

Оскільки нормоване відхилення $t = \frac{x - \bar{x}}{\sigma}$, то форму кривої нормального розподілу визначають двома параматрами: середньою арифметичною і середнім квадратичним відхиленням. Знаючи ці параметри, можна побудувати криву нормального розподілу. При її побудові за емпіричними даними застосовують таку формулу:

$$y = \frac{in}{\sigma} f(t),$$

де y — ордината кривої (теоретична частота); i — інтервал; n — чисельність сукупності; σ — середнє квадратичне відхилення; $f(t)$ — функція щільності нормального розподілу.

Розподіл *t*-Стьюдента — це закон розподілу нормованого відхилення при малих вибірках (n ‹ 20). Цей розподіл симетричний відносно середньої, але, на відміну від нормального, площа, яка знаходиться між віссю абсцис і кривою, не постійна, а залежить від обсягу вибірки. Чим менша кількість ступенів свободи (n – 1), тим більше розтягуються вітки кривої. При великому обсязі вибірки криві розподілу *t*-Стьюдента і нормального розподілу збігаються і вся сукупність знаходиться в межах від $-3t$ до $+3t$. При n = 5 на осі абсцис укладається ± 4,5t, при n = 1 — більше ± 5t.

Розподіл t-Стьюдента описується рівнянням

$$y = y_o \frac{1}{\left(1 + \frac{t^2}{n-1}\right)^{\frac{n}{2}}},$$

де y — ймовірність певного відхилення; y_o — максимальна ймовірність; t — нормоване відхилення = $\frac{\tilde{x} - \bar{x}}{\sigma : n}$.

Розподіл *t*-Стьюдента не залежить від середньої величини та, дисперсії генеральної сукупності, а залежить тільки від обсягу вибірки. Його використовують переважно для оцінки параметрів малої вибірки, коли генеральна сукупність має нормальний розподіл, а також для побудови інтервальних оцінок, перевірки статистичних гіпотез.

Розподіл χ^2 (хі-квадрат) — це закон розподілу вибіркової дисперсії параметрів, які підпорядковуються закону нормального розподілу при малих вибірках. Крива, що характеризує розподіл χ2, описується рівнянням

$$y = y_0\left(\chi^2\right) \cdot \frac{\upsilon - 2}{2} e^{-\frac{1}{2}t^2},$$

де y — ймовірність; y_0 — максимальна ймовірність; υ — кількість ступенів свободи варіації; e — основа натуральних логарифмів.

Розподіл χ^2 не залежить від середньої генеральної та дисперсії. Його визначають кількістю ступенів свободи варіації. При великій чисельності вибірки він набуває форми, близької до нормального розподілу. Розподіл χ^2 для випадкових величин складається з великої кількості розподілів.

За допомогою критерію χ^2 оцінюють незалежність у розподілі об'єктів сукупності за градаціями досліджуваної ознаки, згоду (відповідність) між фактичним і теоретично очікуваним розподілами, а також однорідність розподілів. Відповідно до цього критерій χ^2 називають ще **критерієм незалежності, згоди і однорідності**.

До критерію χ^2 складено спеціальні таблиці, в яких наведені його значення при певній кількості ступенів свободи та вибраному рівні істотності (дод. 6). Наприклад, при 10 ступенях свободи варіації і рівні істотності 0,05 χ^2 = 18,307. Це означає, що внаслідок дії випадкових факторів значення $\chi 2$ може дорівнювати або бути більш як 18,307 у 5 випадках із 100. Якщо фактичне значення χ^2, обчислене за даними вибірки, буде менше за табличне при відповідній кількості ступенів свободи варіації і рівні істотності, то це означає, що розбіжність між фактичним і теоретичним розподілом випадкова. Коли ж фактичне значення χ^2 більше за табличне, то це свідчить, що розбіжності між фактичним і теоретичним розподілом не випадкові.

Обов'язковою умовою застосування критерію χ^2 є достатня чисельність спостережень як у цілому, так і по окремих групах. Обсяг вибірки має бути не менш як 50 одиниць. У кожній групі повинно бути не менше одного спостереження, а груп з частотою менше 5 одиниць — не більш як 20 % обсягу вибірки.

***F*-розподіл** — це спільний закон розподілу двох взаємопов'язаних вибіркових дисперсій для випадкових величин x і y, кожна з яких розподілена нормально.

Щоб висновки про параметри генеральної сукупності, зроблені на основі вибіркового спостереження, були обгрунтованими та надійними, потрібно правильно використовувати методи оцінки параметрів розподілу. При великих вибірках усі вибіркові характеристики підпорядковуються тому самому закону розподілу, що й генеральні характеристики. Так, середня арифметична підпорядковується закону нормального розподілу, причому вона є незміщеною, надійною, ефективною і достатньою оцінкою. У зв'язку із цим уважається, що генеральна середня дорівнює вибірковій середній. Вибіркова дисперсія підпорядковується закону розподілу χ^2 Пірсона, який із збільшенням обсягу вибірки наближається до нормального. Ця дисперсія є зміщеною оцінкою, але при досить великих вибірках поправку Бесселя можна не враховувати. У зв'язку із цим уважається, що генеральна дисперсія дорівнює вибірковій.

3.3 Способи формування вибіркових сукупностей і статистична оцінка вибіркових характеристик

Важливою умовою наукової організації вибіркового спостереження є правильне формування вибіркової сукупності. Для одержання вірогідних даних вибірка повинна достатньо точно відображувати основні характеристики генеральної сукупності.

Вибіркову сукупність, яка достатньо точно відображує генеральну сукупність, називають **репрезентативною**. Щоб вибірка правильно відображувала генеральну сукупність, вона повинна охоплювати потрібну кількість одиниць спостереження, а процес формування вибірки слід організувати так, щоб забезпечити однакову можливість кожній одиниці генеральної сукупності потрапити у вибіркову сукупність.

За способом відбору одиниць для спостереження розрізняють такі види формування вибіркової сукупності: власне випадкову, механічну, серійну і типову вибірки.

Власне випадковою називають таку вибірку, при якій кожна одиниця з генеральної сукупності відбирається у вибірку випадково, ненавмисно. При цьому генеральна сукупність не розподіляється на складові частини. Відбір одиниць звичайно проводиться жеребкуванням.

На практиці застосовують два різновиди власне випадкової вибірки: повторний і безповторний. При **повторній вибірці** кожна раніше відібрана одиниця знову повертається в генеральну сукупність і може знову брати участь у вибірці. Цей вид вибірки забезпечує незалежність наступних витягів від попередніх, оскільки склад генеральної сукупності незмінний.

При **безповторній вибірці** кожна раніше відібрана одиниця не повертається в генеральну сукупність і у дальшому відборі не бере участі. Отже, при повторній вибірці окрема одиниця сукупності може потрапити у вибірку кілька разів, а при безповторній — тільки

один раз. Безповторна вибірка, як правило, дає точніші результати, ніж повторна.

Результати власне випадкової вибірки оцінюють за такими формулами:

- при визначенні середнього розміру ознаки

$$\Delta x = t\sqrt{\frac{\sigma_0^2}{n}} - \text{повторна вибірка;}$$

$$\Delta x = t\sqrt{\frac{\sigma_0^2}{n}\left(1-\frac{n}{N}\right)} - \text{безповторна вибірка;}$$

- при визначенні частки ознаки

$$\Delta p = t\sqrt{\frac{\omega(1-\omega)}{n}} - \text{повторна вибірка;}$$

$$\Delta p = \sqrt{\frac{\omega(1-\omega)}{n}\left(1-\frac{n}{N}\right)} - \text{безповторна вибірка.}$$

Як зазначалося, при проведенні вибіркового спостереження даних про розмір дисперсії (середнього квадрата відхилень) у генеральній сукупності немає. У зв'язку з цим замість них використовують вибіркову дисперсію, скориговану при невеликій кількості спостережень на коефіцієнт $\frac{n}{n-1}$.

Формула помилки при безповторній вибірці відрізняється від відповідної формули при повторній вибірці тільки множником $1-\frac{n}{N}$. Це зумовлене тим, що при безповторній вибірці відібрані одиниці не повертаються у генеральну сукупність і її чисельність

поступово зменшується. Якщо процент вибірки невеликий, то відношення $\frac{n}{N}$ є невелике число, а різниця $1-\frac{n}{N}$ мало чим відрізняється від одиниці. У такому разі цей множник можна не враховувати і граничну помилку вибірки визначати за формулами повторної вибірки.

Розглянемо порядок обчислення граничних помилок вибіркового спостереження на такому прикладі. У спеціалізованому господарстві з вирощування нетелей для визначення приросту живої маси ремонтних телиць у віці від 12 до 15 місяців проведено контрольне зважування. З групи 500 голів взято за схемою власне випадкової безповторної вибірки 50 голів. Результати контрольного зважування телиць наведено в табл. 26.

Таблиця 26 *Розподіл ремонтних телиць контрольної групи за живою масою*

Жива маса телиці, кг	Кількість телиць
270 – 280	6
280 – 290	11
290 – 300	19
300 – 310	10
310 – 320	4

З імовірністю p=0,95 потрібно визначити середню живу масу ремонтних телиць у генеральній сукупності і частку телиць живою масою 300 кг і більше.

Спочатку визначимо середню живу масу телиць у вібірковій сукупності:

$$\bar{x}=\frac{\sum xf}{\sum f}=\frac{275\cdot 6+285\cdot 11+285\cdot 19+305\cdot 10+315\cdot 4}{6+11+19+10+4}=294 \ кг.$$

Вибіркова дисперсія живої маси телиць

$$\sigma_0^2=\frac{\sum(x-\bar{x})^2 f}{\sum f}=\frac{19^2\cdot 6+9^2\cdot 11+1^2\cdot 19+11^2\cdot 10+21^2\cdot 4}{6+11+19+10+4}=121.$$

Скоригуємо вибіркову дисперсію на множник $\frac{n}{n-1}$:

$$\sigma_0^2=121\cdot\frac{50}{50-1}=123{,}47.$$

За таблицею «Значення інтегралу ймовірності» (дод. 1) встановимо, що при рівні ймовірності 0,95 нормоване відхилення (коефіцієнт надійності) — t = 1,96.

Отже, у нашому прикладі $\bar{x}=294$; $\sigma_0^2=123{,}47$; $t=1{,}96$; $n=50$, N = 500.

Підставивши ці дані у формулу граничної помилки вибірки при безповторній вибірці, матимемо:

$$\Delta x=t\sqrt{\frac{\sigma_0^2}{n}\left(1-\frac{n}{N}\right)}=1{,}96\sqrt{\frac{123{,}47}{50}\left(1-\frac{50}{500}\right)}=2{,}9кг.$$

Оскільки помилка вибірки може мати додатний або від'ємний знак, то можливі межі середньої величини у генеральній сукупності визначають за формулою

$$\bar{x}=\bar{\bar{x}}\pm\Delta x \text{ або } \bar{\bar{x}}-\Delta x\le\bar{x}\le\bar{\bar{x}}+\Delta x;$$
$$294-2{,}9\le 294+2{,}9;$$
$$291{,}1\le 296{,}9.$$

Отже, з імовірністю 0,95 можна гарантувати, що різниця між вибірковою середньою та генеральною середньою не перевищить

2.9 кг, а середня жива маса всіх телиць знаходиться в межах від 291,1 до 296,9 кг.

Частка телиць живою масою 300 кг і більше у вибірковій сукупності (частість) становитиме

$$\omega = \frac{m}{n} = \frac{14}{50} = 0{,}28.$$

Скоригована дисперсія альтернативної ознаки

$$\omega = (1-\omega) = 0{,}28 \cdot 0{,}72 \cdot \frac{50}{50-1} = 0{,}206.$$

Гранична помилка частки телиць з живою масою 300 кг і більше

$$\Delta p = t\sqrt{\frac{\omega(1-\omega)}{n}\left(1-\frac{n}{N}\right)} = 1{,}96\sqrt{\frac{0{,}206}{50}\left(1-\frac{50}{500}\right)} = 0{,}061, \text{ або } 6{,}1\ \%.$$

Можливі межі частки у генеральній сукупності обчислюють за формулою

$$p = \omega \pm \Delta_p, або \omega - \Delta_p \le p \le \omega + \Delta_p;$$

$$28{,}0 - 6{,}1 \le 29{,}0 + 6{,}1;$$

$$21{,}0 \le 34{,}1.$$

Отже, з імовірністю 0,95 можна гарантувати, що різниця між частістю і часткою не перевищить 0,061, а частка телиць живою масою 300 кг і більше у генеральній сукупності лежить в межах від 21,9 до 34,1 %.

Механічна вибірка є різновидом випадкової вибірки. При ній одиниці для вибіркового спостереження відбирають не жеребкуванням, а механічно через відповідний інтервал. Для цього всі одиниці

генеральної сукупності розподіляють у певному порядку, але так, щоб порядок не був пов'язаний з розміром досліджуваної ознаки. Наприклад, сукупність населених пунктів можна розташувати за географічним положенням, працівників — за алфавітом прізвищ, однойменні товари — в міру їх виробництва чи надходження тощо.

Якщо загальна чисельність генеральної сукупності 2000 одиниць, а потрібно сформувати вибірку з 200 одиниць, то для вибіркового спостереження відбирають кожну десяту одиницю

$$\frac{N}{n} = \frac{2000}{200} = 10.$$

Розподіливши одиниці генеральної сукупності у певному порядку (за алфавітом, у зростаючому чи спадному порядку) навмисно або за допомогою жеребкування в першій десятці визначають початкову одиницю відбору, наприклад 5. Після цього для вибіркового спостереження відбирають 5,15,25-ту і т.д. одиниці.

Механічну вибірку застосовують при відборі домогосподарств для бюджетних обстежень, при відборі населених пунктів для контрольних обходів з метою уточнення результатів обліку худоби, при контролі якості продукції і т.д. Недоліком цього виду відбору є те, що для його застосування потрібно мати повний облік усіх одиниць генеральної сукупності.

Механічна вибірка завжди безповторна. Оцінку середньої (при $t = 1$) і граничної помилок здійснюють за тими самими формулами, що й при власне випадковій вибірці.

При **серійній (гніздовій) вибірці** для спостереження відбирають не окремі одиниці генеральної сукупності, а серії (гнізда) таких одиниць. Відбір серій проводять власне випадковим або механічним способом. У відібраних серіях обчислюють усі одиниці без винятку.

Загальне число серій, що становлять генеральну сукупність, розглядається при серійній вибірці як її загальна чисельність N_c, а кількість відібраних серій — як чисельність вибірки n_c. Визначаючи помилки серійної вибірки, враховують тільки варіацію ознаки між

окремими серіями, так звану міжсерійну варіацію. Результати серійної вибірки оцінюють за такими формулами:

При визначенні середнього розміру ознаки

$$\Delta_x = t\sqrt{\frac{\sigma_c^2}{n_c}} \text{ — повторна вибірка;}$$

$$\Delta_x = t\sqrt{\frac{\sigma_c^2}{n_c}\left(1-\frac{n_c}{N_c}\right)} \text{ — безповторна вибірка;}$$

при визначенні частки ознаки

$$\Delta p = t\sqrt{\frac{\omega_c(1-\omega_c)}{n_c}} \text{ — повторна вибірка;}$$

$$\Delta p = t\sqrt{\frac{\omega_c(1-\omega_c)}{n_c}\left(1-\frac{n_c}{N_c}\right)} \text{ — безповторна вибірка.}$$

де σ_c^2 — міжсерійна (міжгрупова) дисперсія; ω_c — частість ознаки в середньому у всіх обстежених серіях; N_c — загальна чисельність рівновеликих серій у генеральній сукупності; n_c — чисельність серій, відібраних для обстеження.

Переваги серійної вибірки в тому, що відбирати й обстежувати групи одиниць значно простіше, ніж окремі одиниці. Проте в зв'язку з тим, що при цьому відборі порушується рівномірність розподілу одиниць вибіркової сукупності у всій сукупності, серійна вибірка, як правило, дає більш високу помилку вибірки. Для того щоб забезпечити потрібну точність вибірки, слід збільшувати, порівняно з іншими видами відбору, її чисельність. Прикладом серійної вибірки можуть бути контрольні 10 % обходи, які проводять після перепису худоби в господарствах населення.

При **типовій (районованій) вибірці** всю генеральну сукупність після попереднього аналізу розподіляють на однорідні типові групи, райони, зони за певними ознаками. Потім з кожної групи у випадковому або механічному порядку відбирають певну кількість одиниць у вибіркову сукупність. При цьому кількість відібраних з кожної групи одиниць пропорційна чисельності груп або середнім квадратичним відхиленням усередині типових груп.

Типова вибірка дає точний результат порівняно з іншими видами відбору, тому що розподіл генеральної сукупності на типові групи забезпечує попадання у вибірку одиниць, які належать до різних типових груп.

Репрезентативність типової вибірки залежить від того, наскільки точно віодбражують кожну типову групу відібрані одиниці. Точність типової вибірки для всієї сукупності залежить від варіації ознаки всередині окремих груп, оскільки загальна варіація при розподілі сукупності на типові групи зменшується на величину міжгрупової варіації. Через це при визначенні помилки типової вибірки для середнього розміру ознаки беруть не загальну дисперсію, а середню із часткових (групових) дисперсій. Так само при визначенні помилки вибірки для частки ознаки замість добутку $\omega(1-\omega)$ беруть середню з групових добутків $\omega(1-\omega)$.

Результати типової вибірки оцііюють за такими формулами:

- при визначенні середнього розміру ознаки

$$\Delta_x = t\sqrt{\frac{\overline{\sigma}_0^2}{n}} \text{ — повторна вибірка;}$$

$$\Delta_x = t\sqrt{\frac{\overline{\sigma}_0^2}{n}\left(1-\frac{n}{N}\right)} \text{ — безповторна вибірка;}$$

- при визначенні частки ознаки

$$\Delta p = t\sqrt{\frac{\overline{\omega(1-\omega)}}{n}} \text{ — повторна вибірка;}$$

$$\Delta p = t\sqrt{\frac{\overline{w(1-w)}}{n}\left(1-\frac{n}{N}\right)}$$ — безповторна вибірка;

Обчислення граничної помилки вибіркового спостереження при типовій вибірці розглянемо на такому прикладі. Для обстеження агротехніки вирощування соняшнику в аграрних формуваннях області відібрано 40 господарств. Спочатку всю територію області залежно від природно-кліматичних умов розподілили на 5 типових зон. І кожної зони за принципом випадкової безповторної вибірки відібрали 10 % господарств. Кількість господарств і урожайність соняшнику у відібраних господарствах наведено в табл. 27.

Таблиця 27 *Вихідні та розрахункові дані для визначення помилки типової вибірки*

Зона	Кількість господарств у генеральній сукупності	Вибіркова сукупність			Сума квадратів	Квадрат суми	Середній квадрат суми
		кількість господарств	сума урожайності	середня урожайність, ц/га			
	N_i	n_i	$\sum x$	$\tilde{x}$	$\sum x^2$	$(\sum x)^2$	$(\sum x)^2 : n_i$
Північна	71	7	112,7	16,1	.1818,93	12701,29	1814,47
Східна	80	8	140,0	17,5	2451,50	19600,00	2450,00
Ценральна	92	9	165,6	18,4	3065,16	27423,36	3047,04
Західна	88	9	170,1	18,9	3223,61	28934,01	3214,89
Південна	69	7	145,6	20,8	3035,16	21199,36	3028,48
Разом	400	40	734,0	18,4	13594,36	—	13554,88

Для того, щоб визначити залишкову дисперсію, обчислимо спочатку загальну і міжгрупову дисперсії.

Загальна дисперсія

$$\sigma_3^2 = \frac{\sum x^2}{n} - \bar{x}^2 = 339{,}96 - 338{,}56 = 1{,}30;$$

міжгрупова дисперсія

$$\sigma_м^2 = \frac{\sum \frac{(\sum x)^2}{n_i}}{n} - \bar{x}^2 = 338{,}87 - 338{,}56 = 0{,}31.$$

Середня залишкова дисперсія становить різницю між загальною і міжгруповою дисперсіями:

$$\overline{\sigma_{зал.}^2} = \sigma_3^2 - \sigma_м^2 = 1{,}30 - 0{,}31 = 0{,}99.$$

Скоригуємо вибіркову залишкову дисперсію на множник : $\frac{n}{n-1}$

$$\overline{\sigma_{зал}^2} = 0{,}99 \cdot \frac{40}{39} = 1{,}02.$$

Візьмемо рівень ймовірності p = 0,95 (нормоване відхилення = 1,96) і обчислимо граничну помилку при безповторній вибірці:

$$\Delta_x = t\sqrt{\frac{\overline{\sigma^2}}{n}\left(1-\frac{n}{N}\right)} = 1{,}96\sqrt{\frac{1{,}02}{40}\left(1-\frac{40}{400}\right)} = 0{,}3 ц / га.$$

Отже, з імовірністю помилитися лише в 5 випадках із 100 можна стверджувати, що середня урожайність соняшнику у генеральній сукупності знаходиться в межах $\bar{x} = \tilde{x} \pm \Delta_x = 18{,}4 \pm 0{,}3 = 18{,}1 \le 18{,}7 ц/га$.

Типову вибірку часто проводять кількома стадіями (ступенями). При цьому кожна стадія має свою одиницю відбору. Така вибірка називається **багатоступінчастою**. Особливістю багатоступінчастої вибірки є те, що спочатку з генеральної сукупності відбирають частину одиниць, а потім з цієї частини формують вибірку другого порядку, яку й аналізують. За двоступінчастою вибіркою здійснюють аналіз якості насіння. Спочатку відбирають проби з партії насіння, а потім з відібраного насіння виділяють наважку для визначення схожості, чистоти й інших посівних якостей насіння. За двоступінчастою схемою формують вибірки для аналізу якості продукції. Прикладом триступінчастої вибірки може бути обстеження особистих господарств населення, при якому спочатку відбирають адміністративні райони в окремих регіонах країни, потім сільські (міські) ради, а в них — господарства.

При проведенні вибіркового спостереження важливо знати, наскільки точно вибіркова сукупність відображує (репрезентує) генеральну сукупність. Для цього порівнюють відомі та найбільш важливі показники генеральної сукупності (середнє значення ознаки, частку, дисперсію, середнє квадратичне відхилення) з відповідними показниками вибіркової сукупності. Показник репрезентативності вибірки визначають як відношення вибіркової характеристики до відповідної характеристики генеральної сукупності. Якщо урожайність соняшнику в господарствах, відібраних для вибіркового спостереження, становить 18,4 ц/га, а у всіх господарствах області — 19,2 ц/га, то показник репрезентативності

$$p_n = \frac{\bar{x}_0}{x} = \frac{18{,}4}{19{,}2} = 0{,}958, \text{ або } 95{,}8\ \%.$$

Отже, показник вибіркової врожайності дещо менший генеральної, але ця розбіжність незначна — 4,2 %. На практиці відбір уважають задовільним, якщо показник репрезентативності відбору лежить в межах від 95 до 105 %.

Якщо показник репрезентативності виходитиме за ці межі, то відбір вважається незадовільним і його потрібно повторити. Якщо і повторний відбір буде незадовільним, то слід збільшити чисельність вибірки.

Коли показники вибірки досить точно репрезентують генеральну сукупність, можна досконало вивчити тільки вибіркову сукупність, уважаючи, що знайдені дані вірогідно характеризуватимуть сукупність у цілому.

Заключним етапом вибіркового спостереження є поширення його результатів на всю генеральну сукупність, тобто визначення генеральних показників за вибірковими даними. Розрізняють два способи такого поширення: спосіб прямого перерахування і спосіб коефіцієнтів.

При **прямому перерахуванні** вибіркову середню ознаки або частість множать на чисельність одиниць генеральної сукупності. Припустимо, що середньорічний надій молока від однієї корови для вибіркової сукупності становить 4000 кг, а середньорічна кількість корів в особистих господарствах населення району — 2000 голів. Звідси валовий надій молока в особистих господарствах населення дорівнюватиме 40 · 2000 = 80 тис.ц. Якщо при цьому відомо, що гранична помилка вибірки з імовірністю p = 0,95 дорівнює ±120 кг, то середньорічний надій молока від корови у генеральній сукупності коливатиметься від 3880 до 4120 кг, а валовий надій молока — від 77,6 до 82,4 тис. ц.

Спосіб коефіцієнтів застосовують для уточнення даних суцільного спостереження. При цьому, порівнючи дані вибіркового спостереження з даними суцільного, обчислюють поправочний коефіцієнт, яким і користуються для внесення поправок у матеріали суцільного спостереження. Так, за даними суцільного обліку в особистих

господарствах населення було зареєстровано 1910 корів. Контрольними обходами було охоплено 10 % дворів, у яких зареєстровано 210 корів, а за даними суцільного обліку налічувалося 200 корів. У цьому разі поправочний коефіцієнт дорівнюватиме 210 : 200 = 1,05, а фактичне поголів'я корів в особистих господарствах населення з поправкою на недооблік — 1910 · 1,05 = 2005.

3.4 Визначення потрібної чисельності вибірки

При організації вибіркового спостереження, особливо коли воно проводиться вперше, важливо правильно визначити чисельність вибіркової сукупності. Якщо спостереженню підлягає недостатня кількість одиниць, то знайдені результати будуть неточними і тоді можна зробити необгрунтовані висновки про середній розмір або частку ознаки у генеральній сукупності. Якщо відбирається занадто велика кількість одиниць, то це призводить до зайвих затрат праці та коштів, а при контролі за якістю продукції (схожості насіння, жирності молока) — і до зайвих втрат.

Чисельність вибіркової сукупності залежить від способу відбору одиниць для спостереження, рівня варіації досліджуваної ознаки, розміру граничної помилки вибірки, а також рівня ймовірності, з якою потрібно гарантувати результати вибіркового спостереження. Формули для визначення потрібної чисельності вибірки виводяться з формул граничних помилок, а саме при визначенні середнього розміру ознаки:

$$n = \frac{t^2\sigma_0^2}{\Delta_x^2}$$ — власне випадкова повторна вибірка;

$$n = \frac{t^2\sigma_0^2 N}{\Delta_x^2 N + t^2\sigma_0^2}$$ — власне випадкова та механічна безповторна вибірка;

$$n = \frac{t^2\omega_c(1-\omega_c)N_c}{\Delta_p^2 N_c + t^2\omega_c(1-\omega_c)}$$ — серійна безповторна вибірка;

$$n = \frac{t^2\overline{\omega(1-\omega)}N}{\Delta_p^2 N + t^2\omega(1-\omega)}$$ — типова безповторна вибірка.

При практичному застосуванні формул для визначення обсягу вибірки також немає даних про варіацію досліджуваної ознаки в

генеральній сукупності. При цьому, якщо у формулі для визначення помилки вибірки генеральну дисперсію можна замінити скоригованою випадковою дисперсією, то у формулі для визначення чисельності вибірки цього зробити не можна, оскільки обсяг вибірки слід визначати до початку проведення вибіркового спостереження. Для цього замість фактичних даних генеральної дисперсії беруть дані попередніх спостережень, або проводять пробні обстеження і на їх основі визначають орієнтовні розміри дисперсії. Наприклад, перед тим як приступити до визначення втрат під час збирання урожаю гороху, потрібно визначити кількість метрівок (квадратних рамок площею 1 м^2), які слід накласти на полі площею 100 га. При цьому ставиться умова, щоб з імовірністю 0,95 помилка не перевищила 0,3 ц/га. Для визначення очікуваної варіації втрат урожаю під час збирання гороху проведено пробне обстеження, під час якого встановлено, що дисперсія втрат дорівнює 4 г/м^2 (0,4 ц/га). Метрівки накладають по діагоналі за способом механічної безповторної вибірки.

Звідси потрібну чисельність вибірки визначають за формулою

$$n = \frac{t^2\sigma_0^2 N}{\Delta_x^2 N + t^2\sigma_0^2}.$$

Оскільки чисельність генеральної сукупності N дорівнює 100 га або 1млн. м^2, тобто є настільки великою, що розрахунки можна вести за формулою для повторної вибірки, то

$$n = \frac{t^2\sigma_0^2}{\Delta_x^2} = \frac{1{,}96^2 \cdot 0{,}4}{0{,}3^2} = 17 \text{ метрівок.}$$

Отже, з імовірністю помилитися лише в 5 випадках із 100 можна стверджувати, що коли буде накладено 17 метрівок, то втрати під час збирання гороху будуть визначені з точністю ±0,3 ц/га.

Якщо вибірку здійснюють з невеликої за чисельністю генеральної сукупності, її обсяг визначають за формулою для безповторної вибірки. Наприклад, потрібно розрахувати чисельність вибірки для визначення середньої живої маси поросят при відлученні у селянському господарстві з помилкою не більш як 1 кг. Середнє квадратичне відхилення живої маси поросят при відлученні в господарстві дорівнює близько 2 кг, загальна кількість поросят $N = 40$, рівень ймовірності $p = 0{,}95$. Звідси

$$n = \frac{t^2\sigma_0^2 N}{\Delta_x^2 N + t^2\sigma_0^2} = \frac{1{,}96^2 \cdot 2^2 \cdot 40}{1^2 \cdot 40 + 1{,}96^2 \cdot 2^2} = 11.$$

Отже, з імовірністю 0,95 можна гарантувати, що коли буде відібрано для обстеження у випадковому порядку 11 голів, то жива маса поросят при відлученні буде визначена з точністю ±1 кг.

Для частки ознаки в сукупності дисперсію визначають так: беруть максимальне значення дисперсії альтернативної ознаки 0,25 (якщо $p = 0{,}5$, то $p(1-p) = 0{,}5 \cdot 0{,}5 = 0{,}25$).

При визначенні потрібної чисельності вибірки помилку часто обчислюють у процентах. У цьому разі варіювання досліджуваної ознаки слід також виразити у відносних величинах і дисперсію замінити коефіцієнтом варіації.

Коли вибіркове спостереження проводять для визначення середніх і відносних показників кількох ознак, потрібно встановити таку чисельність вибірки, яка забезпечила б задовільні результати за всіма ознаками.

3.5 Мала вибірка

Щоб вибіркова сукупність правильно відображувала генеральну сукупність, вона повинна охоплювати достатню кількість одиниць спостереження. Чим більше одиниць спостереження включає вибірка, тим правильнішим буде висновок про розмір ознаки у генеральній сукупності. Однак на практиці не завжди доцільно одержувати великі вибіркові сукупності. Так, при перевірці якості продукції в дослідній справі обмежуються порівняно невеликими за обсягом вибірковими сукупностями.

Вибірки, чисельність яких не перевищує 20 одиниць спостереження, називають **малими вибірками**. Невеликий обсяг малої вибірки деякою мірою знижує її точність порівняно із звичайною вибіркою, чисельність якої перевищує 20 одиниць спостереження. Проте математична статистика розробила способи, які дають змогу вірогідно оцінювати результати малої вибірки і поширювати їх на генеральну сукупність. При цьому розрахунок середньої та граничної помилок має деякі особливості.

При великому обсязі вибіркової сукупності ($n > 20$) співвідношення між генеральною та вибірковою дисперсіями матиме такий вигляд:

$$\sigma^2_{ген} = \sigma^2_{виб}\left(\frac{n}{n-1}\right).$$

Якщо вибіркова сукупність досить велика, то множник $\frac{n}{n-1}$ наближається до одиниці. Так, при $n = 100$ він дорівнює 1,01, при $n = 200$ – 1,005, при $n = 500$ – 1,002 і вибіркова дисперсія збігається з генеральною. У невеликих за чисельністю сукупностях ($n<20$) цей множник слід обов'язково враховувати, тому дисперсію в малих вибірках обчислюють за такою формулою:

$$\sigma_{м.в.} = \frac{\sum(x-\bar{x})^2}{n-1},$$

де $n - 1$ — кількість ступенів свободи варіації.

Під **ступенем свободи варіації** розуміють кількість варіантів, які можуть мати довільні значення, не змінюючи їх загальної характеристики (середньої). Наприклад, є дані трьох спостережень: $x_1 = 9$, $x_2 = 11$, $x_3 = 16$. Звідси

$$\bar{x} = \frac{9+11+16}{3} = 12.$$

При трьох спостереженнях вільно варіюючих величин залишається тільки дві, тому що третю можна визначити з відомих двох величин і середньої. Якщо відомі величини x_1 і x_2, то величину x_3 визначають як різницю $nx - (x_1 + x_2)$. У нашому прикладі 36 – 9 – – 11 = 16.

Результати малої вибірки за допомогою критеріїв стандартної нормальної кривої (як при звичайній вибірці) оцінити не можна, оскільки при невеликій кількості спостережень розподіл імовірностей для середньої значною мірою залежить від характеру розподілу індивідуальних величин.

Основи теорії малої вибірки розробив англійський математик-статистик Вільям Госсет (псевдонім Стьюдент). Дослідження В. Госсета показали, що при невеликій кількості спостережень середнє квадратичне відхилення вибіркової сукупності значно відрізняється від середнього квадратичного відхилення генеральної сукупності. У зв'язку з цим використання кривої нормального розподілу для оцінки даних малої вибірки дає приблизні результати.

В. Госсет обгрунтував закон розподілу відхилень вибіркових середніх від генеральної середньої для малих вибірок. Згідно із цим законом імовірна оцінка того, що гранична помилка не перевищить

t-кратну середню помилку в малих вибірках, залежить не тільки від значення *t*, а й від обсягу вибірки.

Теоретично нормоване відхилення для малих вибірок дістало назву критерію *t*-Стьюдента на відміну від критерію t-нормального розподілу, який застосовують для великих вибірок. Із збільшенням обсягу вибірки розподіл Стьюдента наближається до нормального розподілу.

На основі встановленої закономірності розподілу помилок малих вибірок складено спеціальні таблиці, в яких наведено значення критерію *t*-Стьюдента і відповідних рівнів ймовірності при різній чисельності одиниць вибірки. Ці таблиці бувають двох видів. У таблиці «Значення імовірностей для критерію *t*-Стьюдента» (дод. 4) показано ймовірності певного розміру помилки залежно від величини критерію t і чисельності одиниць вибірки. Наприклад, при $t = 2$ і 5 ступенях свободи рівень ймовірності помилки становить 0,884. Це означає, що у 884 випадках із 1000 визначена помилка вибірки не перевищуватиме встановлених розмірів. В останніх 116 випадках фактична помилка може бути більшою за встановлені розміри.

У таблиці «Значення критерію *t*-Стьюдента при рівні ймовірності 0,10; 0,05 і 0,01» (дод. 2) наведено розмір критерію t при встановленому рівні ймовірності і відповідній кількості ступенів свободи варіації. Наприклад, при 5 ступенях свободи і рівні ймовірності 0,05 теоретичне значення критерію *t* дорівнює 2,5706. Це означає, що тільки в 5 випадках із 100 значення нормованого відхилення t через випадкові причини може перевищити зазначену величину (2,5706), а в інших випадках воно буде меншим або таким, як у таблиці. Іншими словами, табличне значення показує максимальну величину відношення випадкових помилок до їх середньої помилки.

Порівняння таблиць імовірностей нормального розподілу та розподілу Стьюдента показує, що при $n = 20$ і більше вони майже не відрізняються. Так, при $t = 1$ ймовірність дорівнює у великих вибірках 0,683, у малих — 0,670, при $t = 2$ відповідно — 0,954 і 0,940, при $t = 3$ – 0,997 і 0,992.

Малі вибірки дають неточні результати порівняно із звичайними вибірками. Тому їх рідко застосовують для встановлення дійсної величини середнього розміру або частки ознаки в генеральній сукупності. Малі вибірки використовують в основному для оцінки ймовірності розбіжностей між показниками вибіркових сукупностей, наприклад, при порівнянні дослідних даних про урожайність сільськогосподарських культур і продуктивність тварин, визначенні розміру втрат під час збирання урожаю на двох ділянках тощо.

Контрольні питання

1. *Які суть, значення і переваги вибіркового спостереження?*
2. *Які теоретичні основи вибіркового спостереження?*
3. *Що таке середня та гранична помилки вибіркового спостереження?*
4. *Які існують закони розподілу вибіркових характеристик?*
5. *Які існують способи формування вибіркових сукупностей?*
6. *Як здійснюється статистична оцінка вибіркових характеристик?*
7. *Як визначається обсяг вибірки?*
8. *Які існують поширення вибіркових даних на всю генеральну сукупність?*
9. *Що таке мала вибірка та які її особливості?*

Розділ 4
ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

4.1 Поняття про статистичні гіпотези

У практичній і науковій діяльності часто доводиться на підставі результатів обстежень перевіряти різні припущення про характеристики масових явищ. Так, при заміні одного сорту якої-небудь сільськогосподарської культури іншим слід перевірити припущення про те, що інший сорт порівняно з першим має вищу урожайність. При впровадженні нових форм організації праці виникає потреба у перевірці припущення про їх вищу ефективність порівняно з існуючими формами. Перевірку таких припущень на підставі даних вибіркового спостереження називають статистичною перевіркою гіпотез.

Статистична гіпотеза — це припущення відносно параметрів або форми розподілу генеральної сукупності, яке можна перевірити на основі вибірки. У процесі перевірки статистичної гіпотези потрібно визначити, чи узгоджуються дані спостереження з висунутим припущенням. Унаслідок перевірки гіпотеза приймається або відхиляється.

Розрізняють гіпотези основні (нульові або робочі) і альтернативні (конкуруючі). **Основною** називають **гіпотезу**, яка підлягає перевірці. Оскільки здебільшого завдання зводиться до перевірки гіпотези про відсутність реальних розбіжностей (нульова розбіжність) між фактичними та гіпотетичними (теоретично очікуваними) показниками, то її називають нульовою і позначають через H_0. **Альтернативною** На називають **гіпотезу**, яка протиставляється нульовій гіпотезі та заперечує її. Кожній нульовій гіпотезі можна протиставити нескінченну множину альтернативних. Зміст H_0 і На залежить від завдань, що вирішуються перевіркою статистичних гіпотез і характером вибіркової сукупності. Запис змісту гипотез має такий вигляд:

$$H_0: x = a; H_a: x \neq a.$$

Якщо нульова гіпотеза полягає в припущенні, що середньодобовий приріст поросят на відгодівлі у генеральній сукупності дорівнює 400 г, то альтернативна гіпотеза може полягати в тому, що середньодобовий приріст поросят у генеральній сукупності не дорівнює 400 г. Ці гіпотези записують так:

$$H_0: \bar{x} = 400; H_a: \bar{x} \neq 400.$$

За формою побудови розрізняють гіпотези прості і складні. **Простою** називають гіпотезу, яка стосується тільки одного припущення (наприклад, H_0: $x = 3$), складною — гіпотезу, яка стосується двох і більше припущень (наприклад, Н0: $\bar{x} > 3$). Складна гіпотеза характеризує деяку галузь імовірних значень досліджуваного параметра.

Якщо внаслідок перевірки нульової гіпотези розбіжності між фактичними та гіпотетичними показниками близькі до нуля або лежать в межах допустимих значень, то нульова гіпотеза не відхиляється, а якщо розбіжності лежать у критичній для певного статистичного критерію галузі і несумісні з нею, то нульова гіпотеза відхиляється. Прийняття нульової гіпотези означає, що дані спостереження не суперечать припущенню про відсутність розбіжностей між фактичними і гіпотетичними показниками, але не доводить того, що цих розбіжностей немає. Відхилення гіпотези означає, що емпіричні дані несумісні з нульовою гіпотезою, а правильна інша, альтернативна, гіпотеза.

Перевірка статистичних гіпотез пов'язана з можливим прийняттям неправильних рішень, тобто з можливістю допущення помилок у висновках. Розрізняють помилки першого та другого порядку. **Помилки першого порядку** полягають у тому, що відхиляється нульова гіпотеза, хоч насправді вона правильна. **Помилки другого порядку** полягають у тому, що приймається нульова гіпотеза, хоча насправді правильною є альтернативна гіпотеза. Правильні та неправильні рішення можна дістати в обох випадках (табл. 28).

Таблиця 28 *Можливі результати перевірки гіпотез*

Результати перевірки нульової гіпотези	Оцінка прийнятого рішення	
	Правильна нульова гіпотеза H_0	Правильна альтернативна гіпотеза *На*
Гіпотеза відхиляється	Помилка першого порядку	Правильне рішення
Гіпотеза приймається	Правильне рішення	Помилка другого порядку

Ймовірність допустити помилку першого порядку дістала назву рівня **значущості** і позначається через α. Він становить ту мінімальну ймовірність, починаючи з якої можна визнати подію неможливою, тобто рівень значущості показує міру, з якою ми ризикуємо, відхиляючи нульову гіпотезу.

Рівень значущості визначає дослідник залежно від характеру та важливості вирішуваних завдань за принципом практичної упевненості. Частіше його встановлюють на рівні 0,05 і 0,01, а у разі відповідальних рішень (наприклад, при перевірці дії пестицидів) його підвищують до рівня 0,001. Рівень значущості α = 0,05 показує, що тільки в 5 випадках із 100 є ризик допустити помилку першого порядку, тобто відхилити нульову гіпотезу.

Чим менший рівень значущості, тим менша ймовірність відхилити нульову гіпотезу, якщо вона правильна (тобто допустити помилку першого порпядку), і тим більша ймовірність допустити помилку другого порядку, якщо нульову гіпотезу не відхиляють (тобто вона неправильна). Рівень значущості не вимірює ступеня ризику пов'язаного з прийняттям неправильної гіпотези (помилки другого порядку), він лише контролює помилку першого порядку.

Оскільки помилки першого та другого порядку є конкуруючими, то зменшення ймовірності допустити одну з них зумовлює збільшення ймовірності допустити другу помилку. Тому в кожному випадку слід вибирати компромісне рішення. Єдиним правильним шляхом одночасного зменшення можливих помилок є збільшення обсягу вибірок.

Для перевірки нульової гіпотези використовують спеціальні статистичні критерії. **Статистичний критерій** — це оціночний показник, обчислений на основі фактичних спостережень, відповідно до якого приймають або відхиляють нульову гіпотезу. Побудова критерію зводиться до вибору відповідної функції, що називається **статистикою критерію**.

Залежно від виду перевірюваної гіпотези використовують спеціально розроблені критерії, серед яких найчастіше застосовують *t*-критерій нормального розподілу, *t*-критерій розподілу Стьюдента, *F*-критерій Фішера-Снедекора, критерій χ^2 Пірсона, критерій Колмогорова (λ), критерій Вілкоксона тощо.

Критерій *t* нормального розподілу — це теоретичне нормоване відхилення для великих вибірок. За законом нормального розподілу варіація індивідуальних значень досліджуваної ознаки перебуває в межах $x \pm 3\sigma$ (правило трьох сигм). Числове значення цього критерію залежить від рівня ймовірності. Його визначають за спеціальними таблицями «Значення інтеграла ймовірностей» (дод. 1).

Критерій *t*-Стьюдента використовують для перевірки статистичних гіпотез стосовно середніх при малій вибірці ($n < 20$). Крім того, його застосовують при визначенні надійних інтервалів, інтервально оцінюючи параметри генеральної сукупності. Числове значення критерію залежить від кількості ступенів свободи варіації та рівня ймовірності (див.дод. 2).

Критерій *F*-Фішера-Снедекора використовують для оцінки співвідношення дисперсій при малих вибірках, а також суті ступеня варіації ознак і надійності взаємозв'язку між факторами (дод. 5).

Критерій χ^2 Пірсона використовують тоді, коли потрібно визначити ступінь відмінності фактичного розподілу частот від теоретичного. Крім того, його застосовують для оцінки однорідності розподілів, а також як критерій незалежності в розподілі об'єктів сукупності за градаціями досліджуваної ознаки (дод. 6).

Критерій λ Колмогорова застосовують для наближеної оцінки ймовірності розбіжностей між фактичними і теоретичними розподілами. При цьому не потрібно визначати кількість ступенів свободи варіації, оскільки граничні значення критерію не залежать від кількості спостережень і є стандартними: для рівня істотності 0,05 $\lambda = 1{,}36$; 0,01 $\lambda = 1{,}63$ і 0,001 $\lambda = 1{,}95$. Нульова гіпотеза відхиляється при 5 %-му рівні істотності, якщо фактичне значення $\lambda < 1{,}36$, при 1 % $\lambda < 1{,}63$ і при 0,1 % $\lambda < 1{,}95$.

Критерій Вілкоксона застосовують для перевірки однорідності розподілів двох генеральних сукупностей. Якщо вибірки здійснено з однієї генеральної сукупності, то додатні і від'ємні різниці між варіантами взаємно погашаються, а їх сума має дорівнювати нулю. Ще з більшою ймовірністю таке припущення можна зробити, якщо розглядати не різниці, а їх ранги. Відхилення суми рангів (додатної або від'ємної) від нуля у цьому разі зумовлене випадковими причинами. Коли ж фактичне відхилення суми рангів від нуля перевищує можливе випадкове відхилення, то нульова гіпотеза відхиляється, тобто відхиляється припущення, що вибірки належать до однієї генеральної сукупності. А якщо це так, то приймається альтернативне припущення, що вибірки зроблені з різних сукупностей і між ними є розбіжності. Критичні значення критерію Вілкоксона наведено в дод. 8. Якщо фактична сума рангів менша за табличне значення або дорівнює йому, то нульова гіпотеза відхиляється на відповідному рівні істотності.

Статистичні критерії, які використовують для перевірки статистичних гіпотез, поділяють на параметричні та непараметричні.

Параметричними називають критерії, які грунтуються на припущенні, що розподіл досліджуваної ознаки в сукупності підпорядковується певному відомому закону, наприклад, закону нормального розподілу, Єтьюдента, Фішера і т.д. До них належать критерії *t*, *F*, χ^2. Особливістю цих критеріїв є те, що їх застосування потребує обчислення оцінок параметрів розподілу.

Непараметричними називають критерії, використання яких не пов'язане із знанням закону розподілу випадкової величини. Їх можна використовувати і тоді, коли досліджуваний розподіл відрізняється від нормального. До непараметричних належать критерії Колмогорова, Вілкоксона, Уайта тощо.

За допомогою непараметричних критеріїв можна швидко дістати оцінки вибіркових спостережень, але вони є менш ефективними порівняно з параметричними критеріями. Тому, якщо розподіл є нормальним або близьким до нормального, слід використовувати параметричні критерії. Якщо ж відомо, що розподіл сукупності, з якої одержана вибірка, відрізняється від нормального, непараметричні критерії дають повнішу інформацію. Єдиною умовою їх застосування є взаємна незалежність даних спостереження, що досягається випадковим відбором.

Під **потужністю критерію** розуміють ймовірність відхилення нульової гіпотези, що перевіряється, коли правильною є альтернативна гіпотеза. Отже, потужність критерію — це ймовірність того, що не буде допущено помилку другого порядку. Із усіх можливих критеріїв потрібно вибирати найбільш потужний. Із збільшенням рівня ймовірності потужність критерію підвищується. При цьому слід урахувати, що цей шлях не завжди прийнятний, оскільки він призводить до необгрунтованого підвищення ймовірності помилок першого порядку.

Під час перевірки гіпотез визначають надійний інтервал відповідного показника. Для цього встановлюють критичну галузь і галузь допустимих значень.

Критична галузь — це ті значення критерію, при яких нульова гіпотеза відхиляється. **Галузь допустимих значень** — це ті значення критерію, при яких нульова гіпотеза приймається. Точки, які відокремлюють критичну галузь від галузі допустимих значень, називають критичними точками.

При перевірці статистичних гіпотез керуються таким правилом: якщо фактичне значення критерію потрапляє у критичну галузь, то нульову гіпотезу відхиляють, а якщо фактичне значення критерію належить до галузі допустимих значень, то нульову гіпотезу приймають.

Отже, при перевірці статистичних гіпотез фактичні значення критерію порівнюють з їх теоретичними значеннями, визначеними за спеціальними таблицями. Якщо фактичне значення критерію, обчислене за даними вибірки, буде більше за табличне значення (при встановленому рівні значущості), то висунуту нульову гіпотезу відхиляють і приймають альтернативну гіпотезу. Коли фактичне значення критерію буде менше за табличне (або дорівнюватиме йому), то роблять висновок про відповідність даних спостереження нульовій гіпотезі і її приймають.

Розрізняють **односторонню** (правосторонню або лівосторонню) і **двосторонню галузі**. Їх вибір при перевірці статистичних гіпотез залежить від конкретних умов і мети вирішуваного завдання. Критичну галузь слід будувати урахуванням відмінності між нульовою гіпотезою та альтернативною.

Двосторонню критичну галузь використовують при інтервальній оцінці параметрів розподілу, односторонню — при вивченні позитивних або негативних відхилень.

Принципова схема перевірки статистичних гіпотез складається з таких послідовно виконуваних етапів:

1) оцінки вихідної інформації й опису статистичної моделі вибіркової сукупності;
2) формулювання нульової й альтернативної гіпотез;

3) встановлення рівня значущості, за якого прийматиметься чи відхилятиметься нульова гіпотеза;
4) вибору найбільш потужного критерію для перевірки нульової гіпотези та розрахунку його фактичного значення;
5) встановлення табличного значення критерію, тобто критичної галузі та галузі прийняття нульової гіпотези;
6) порівняння фактичного і табличного значень критерію і формулювання висновків про прийняття або відхилення нульової гіпотези.

Перевіряючи статистичні гіпотези, найчастіше визначають ймовірність розбіжностей між параметрами статистичних сукупностей (оцінку вірогідності розбіжностей між середніми величинами, дисперсіями, коефіцієнтами регресії і т.д), а також ймовірність відмінностей законів розподілу (оцінку відповідності фактичного розподілу вибіркової сукупності теоретичному, оцінку однорідності складу статистичних сукупностей тощо).

Вибір методу перевірки статистичної гіпотези залежить від обсягу вибіркової сукупності (великі та малі вибірки), чисельності вибірок (рівні або нерівні), рівності дисперсій (рівні або нерівні) і способу формування вибіркових сукупностей.

Залежно від характеру формування розрізняють незалежні та залежні вибірки. **Незалежними** називають вибірки, коли кожному спостереженню в одній вибірці не можна протиставити спостереження другої вибірки. Іншими словами, незалежні вибірки такі, варіанти яких змінюються незалежно один від одного. Прикладом незалежних спостережень можуть бути досліди з двома групами тварин, одна з яких є контрольною, а друга — дослідною і обидві групи сформовані у випадковому порядку.

Коли ж спостереження в одній вибірці якоюсь мірою пов'язані із спостереженнями у другій вибірці, то такі вибірки називають залежними, наприклад, під час досліду тварин спочатку розподіляють на групи за якими-небудь ознаками (масою, віком, статтю і т.д.), а потім з кожної пари аналогів відбирають по одному представнику у

дослідну та контрольну групи. Прикладом залежних вибірок можуть бути також польові досліди з повторностями, кожна з яких розміщується на ділянці порівняно однакової родючості.

4.2 Перевірка статистичних гіпотез про істотність розбіжностей між дисперсіями

У соціально-економічних і сільськогосподарських дослідженнях оцінку істотності розбіжностей між дисперсіями використовують при розв'язанні питань, пов'язаних з вивченням варіації ознак, надійності взаємозв'зку між факторами тощо.

Гіпотезу про істотність розбіжностей між дисперсіями при малих вибірках перевіряють за допомогою критерію *F*-Фішера, який становить відношення двох порівнюваних дисперсій:

$$F = S_1^2 : S_2^2.$$

Критичне значеня критерію *F* знаходять за спеціальними таблицями при встановленому рівні значущості і відповідних кількостях ступенів свободи варіації (дод. 5).

При великих вибірках ($n > 20$) для оцінки істотності розбіжностей між дисперсіями використовують критерій *t*-критерію нормального розподілу:

$$t = \frac{S_1^2 - S_2^2}{\overline{m}_{1-2}},$$

де $\overline{m}_{1-2} = \sqrt{m_1^2 + m_2^2}$; $m_1 = \frac{S_1}{\sqrt{2(n_1 - 1)}}$; $m_2 = \frac{S_2}{2(n_2 - 1)}$.

Критичне значення критерію *t* обчислюють за таблицями (дод. 1).

Поелідовність розрахунків при перевірці гіпотез про існування відмінностей між дисперсіями розглянемо на такому прикладі (табл. 29).

Таблиця 29 *Внутрішньогрупова варіація виробітку родючості ґрунтів у господарствах району*

Групи господарств за родючістю ґрунтів, балів	Кількість господарств	Виробіток робітниць						Загальна сума внутрішньогрупової варіації	Дисперсія	Скориговані дисперсії
		1	2	3	4	5	6			
51 – 55	4	53	53	55	51	—	—	8	2,0	2,67
56 – 60	5	56	56	57	59	57	—	6	1,2	1,50
61 – 65	6	65	64	62	64	65	64	6	1,0	1,20
66 – 70	3	70	66	66	—	—	—	14	4,7	7,05

За даними табл. 29 потрібно перевірити істотність розбіжностей між виутрішньогруповими дисперсіями. Для усунення помилки зміщення скоригуємо обчислені внутрішньогрупові дисперсії на поправочний коефіцієнт Бесселя:

$$S_2 = \sigma^2 \frac{n}{n-1}.$$

Виправлені дисперсії наведено в останній графі таблиці.

Сформулюємо гіпотези: нульову — дисперсії суміжних груп не мають суттєвих розбіжностей; альтернативну — дисперсії суміжних груп суттєво відрізняються:

$$H_0 : S_1^2 = S_2^2; \quad H_a : S_1^2 \neq S_2^2;$$
$$H_0 : S_2^2 = S_3^2; \quad H_a : S_2^2 \neq S_3^2;$$
$$H_0 : S_3^2 = S_4^2; \quad H_a : S_3^2 \neq S_4^2.$$

Установимо рівень значущості $\alpha = 0{,}05$, гарантуючи прийняття гіпотези або відхилення її з імовірністю помилитися тільки у 5 випадках із 100.

Найбільш потужним критерієм для перевірки аналогічних гіпотез при малих вибірках є критерій F-Фішера. Фактичне його значення

$$F_1 = S_1^2 : S_2^2 = 2{,}67 : 1{,}50 = 1{,}78;$$
$$F_2 = S_2^2 : S_3^2 = 1{,}50 : 1{,}20 = 1{,}25.$$
$$F_3 = S_4^2 : S_3^2 = 5{,}55 : 1{,}20 = 4{,}62.$$

Визначимо кількість ступенів свободи варіації для кожної дисперсії:

$$\upsilon_1 = n_1 - 1 = 4 - 1 = 3;$$
$$\upsilon_2 = n_2 - 1 = 5 - 1 = 4;$$
$$\upsilon_3 = n_3 - 1 = 6 - 1 = 5;$$
$$\upsilon_4 = n_1 - 1 = 3 - 1 = 2.$$

Обчислимо табличне значення критерію F (критичну галузь залежно від рівня значущості $\alpha = 0{,}05$ і відповідних ступенів свободи варіації (дод. 5)

$$F_{0,05(1)} = 9{,}19;$$
$$F_{0,05(2)} = 6{,}26;$$
$$F_{0,05(3)} = 19{,}30.$$

Формулюємо правило прийняття рішення за результатами перевірки нульової гіпотези порівнянням фактичного значення критерію з його табличним значенням:

- якщо $F_\phi \geq F_{0,05}$, то фактичний критерій лежить у критичній галузі і висунута нульова гіпотеза відхиляється;
- якщо $F_\phi < F_{0,05}$, то фактичний критерій лежить в галузі допустимих значень і нульова гіпотеза повинна бути прийнята.

Порівнюємо фактичне і табличне значення критерію і робимо висновок про прийняття або відхилення нульової гіпотези:

- $F_1 = 1{,}78 < F_{0,05(1)} = 9{,}19$ — нульова гіпотеза приймається;
- $F_2 = 1{,}25 < F_{0,05(2)} = 6{,}26$ — нульова гіпотеза приймається;
- $F_3 = 4{,}62 < F_{0,05(3)} = 1\ 9{,}30$ — нульова гіпотеза приймається.

Отже, розбіжності між дисперсіями несуттєві. Цей висновок гарантується з імовірністю помилитися в 5 випадках із 100.

4.3 Перевірка статистичних гіпотез відносно середніх величин

Вибір схеми перевірки відносно середніх величин залежить насамперед від чисельності вибірки. При великому обсязі вибірки ($n > 20$) для перевірки гіпотез використовують критерій t-нормального розподілу, при малих вибірках ($n < 20$) — критерій t-Стьюдента.

Схема перевірки гіпотез відносно середніх величин залежить також від рівності вибірок за чисельністю. При рівних ($n_1 = n_2$) і нерівних ($n_1 \neq n_2$) вибірках середню помилку двох вибіркових середніх визначають за різними формулами.

Важливою умовою для визначення схеми перевірки гіпотез є рівність дисперсій. Можливі два варіанти стосовно вибіркових дисперсій: дисперсії рівні ($\sigma_1^2 = \sigma_2^2$) і дисперсії нерівні ($\sigma_1^2 \neq \sigma_2^2$). При суттєвих розбіжностях між дисперсіями фактичне значення t-критерію обчислюють за модифікованими формулами.

На вибір схеми перевірки гіпотез відносно середніх величин істотно впливає спосіб формування вибіркових сукупностей. Якщо дані двох вибірок незалежні, то в цьому разі проводять статистичну оцінку вірогідності різниці середніх, а якщо вибірки залежні, то здійснюють статистичну оцінку вірогідності середньої різниці.

Оцінку вірогідності різниці двох середніх розглянемо на такому прикладі. Для вивчення впливу біоміцину на середньодобові прирости живої маси курчат сформовано контрольну та дослідну групи по 10 голів у кожній. Схема досліду передбачала випадкове формування груп (незалежні вибірки з рівними дисперсіями у генеральних сукупностях). Вихідні та розрахункові дані для оцінки вірогідності різниці середньодобових приростів живої маси курчат наведено в табл.30.

Порівняння середньодобових приростів курчат в обох групах свідчить про те, що більш високі добові прирости дали курчата

дослідної групи. Але у зв'язку з тим, що чисельність вибірок невелика (n = 10), не виключена можливість, що розбіжності між середньодобовими приростами є наслідком дії випадкових причин. Потрібно статистично оцінити різницю середніх в обох групах курчат.

Таблиця 30 *Середньодобові прирости живої маси курчат контрольної і дослідної групи, г*

Номер спостереження	Контрольна група		Дослідна група	
	x_1	x_1^2	x_2	x_2^2
1	20	400	24	576
2	17	289	23	529
3	18	324	22	484
4	19	361	23	529
5	22	484	24	576
6	19	361	25	625
7	21	441	26	676
8	16	256	21	441
9	18	324	20	400
10	22	484	25	625
Разом	192	3724	233	5461

Сформулюємо нульову гіпотезу про те, що добавка до корму дослідної групи курчат біоміцину не впливала на середньодобові прирости живої маси, тобто абсолютна різниця між середніми добовими приростами лежить в межах випадкових коливань:

$$H_0 : \overline{x_2} = \overline{x_1};\ \ H_в : \overline{x_2} \neq \overline{x_1}.$$

Приймемо рівень значущості 0,05, гарантуючи прийняття гіпотези або відмову від неї з імовірністю помилитися лише в 5 випадках із 100.

Послідовність розрахунків така. Спочатку обчислимо середньодобові прирости живої маси курчат у контрольній і дослідій групах:

$$\widetilde{x}_1 = \frac{\sum x_1}{n} = \frac{192}{10} = 19{,}2г; \qquad \widetilde{x}_2 = \frac{\sum x_2}{n} = \frac{233}{10} = 23{,}3\ \ г.$$

Отже, різниця між середніми двох вибірок, вірогідність якої потрібно довести або відкинути, становить

$$\Delta = \widetilde{x}_2 - \widetilde{x}_1 = 23{,}3 - 19{,}2 = 4{,}1.$$

Для визначення середніх помилок у кожній вибірці обчислимо скориговані на втрату ступенів свободи дисперсії:

$$\sigma_1^2 = \frac{\sum x_1^2 - n\overline{x}_1^2}{n-1} = \frac{3724 - 10 \cdot 19{,}2^2}{10-1} = 4{,}18;$$

$$\sigma_2^2 = \frac{\sum x_2^2 - n\overline{x}_2^2}{n-1} = \frac{5461 - 10 \cdot 23{,}3^2}{10-4} = 3{,}57.$$

Для кожної вибірки визначимо квадрати середніх помилок:

$$m_1^2 = \frac{\sigma_1^2}{n_1} = \frac{4{,}18}{10} = 0{,}418; \qquad m_2^2 = \frac{\sigma_2^2}{n_2} = \frac{3{,}57}{10} = 0{,}357.$$

Визначимо середню помилку різниці двох середніх:

$$\overline{m}_{1-2} = \sqrt{m_1^2 + m_2^2} = \sqrt{0{,}418 + 0{,}357} = 0{,}88.$$

Знаючи різницю середніх і середню помилку двох середніх, обчислимо фактичне значення t-критерію:

$$t_\phi = \frac{(\widetilde{x}_2 - \widetilde{x}_1)}{\overline{m}_{1-2}} = \frac{4{,}1}{0{,}88} = 4{,}6591.$$

Кількість ступенів свободи варіації становить:

$$\nu_1 = n_1 - 1 = 10 - 1 = 9; \quad \nu_2 = n_2 - 1 = 10 - 1 = 9;$$
$$\nu_{1-2} = (n_1 - 1) + (n_2 - 1) = 9 + 9 = 18.$$

При рівні значущості 0,05 і 18 ступенях свободи варіації теоретичне значення t-критерію становить 2,1009 (дод. 2).

Оскільки фактичне значення $t_\phi = 4{,}6591$ більше за табличне $t_T = 2{,}1009$, то результати досліду не узгоджуються з гіпотезою про те, що різниця середніх випадкова. Тому можна зробити висновок, що різниця в середньодобових приростах живої маси курчат у дослідній і контрольній групах пов'язана з впливом досліджуваного фактора, тобто добавкою біоміцину.

Висновок про гіпотезу можна зробити й іншим шляхом. Для цього обчислюють граничну помилку двох вибіркових середніх:

$$\Delta = t \cdot \overline{m}_{1-2} = 2{,}1009 \cdot 0{,}88 = 1{,}85 \ г.$$

Гранична помилка показує, що якби умови годівлі курчат дослідної і контрольної груп істотно не відрізнялися, то максимально

можлива різниця між середніми як наслідок дії випадкових факторів, дорівнювала б 1,85 г. Фактична різниця дорівнює 4,1 г, що значно перевищує граничну помилку при рівні значущості 0,05. Це дає змогу зробити висновок, що різниця в середньодобових приростах курчат зумовлена добавкою біоміцину.

Дещо складніше оцінити ймовірність різниці двох вибіркових середніх при різній кількості спостережень, тобто нерівній чисельності вибірок. Припустимо, що при перевірці якості насіння різних зразків озимої пшениці одержано такі дані (табл. 31).

Середня маса 1000 насінин сорту Донецька 46 $\widetilde{x} = 40{,}5$ г, сорту Харківська 11 $\widetilde{x}_2 = 38{,}8$ г. При рівні значущості 0,05 потрібно оцінити ймовірність різниці середніх $\widetilde{x} - \widetilde{x}_2 = 1{,}7$ г.

Таблиця 31 *Маса 1000 насінин озимої пшениці*

Сорт	Номер зразка					
	1	2	3	4	5	6
Донецька 46	42	39	43	40	38	4
Харківська 11	36	40	41	37	40	—

Сформулюємо нульову й альтернативну гіпотези:

$$H_0 : \widetilde{x}_1 = \widetilde{x}_2; \quad H_a : \widetilde{x}_1 \neq \widetilde{x}_2.$$

Для їх перевірки спочатку обчислимо суми квадратів відхилень:

- сорту Донецька 46 $\sum x_1^2 - \overline{n}x_1^2 = 9859 - 9841{,}5 = 17{,}5;$
- сорту Харківська 11 $\sum x_2^2 - \overline{n}x_2^2 = 7546 - 7527{,}2 = 18{,}8.$

Визначимо об'єднану дисперсію двох вибірок за формулою

$$\sigma_0^2 = \frac{\sum(x_1 - \tilde{x}_1)^2 + (x_2 - \tilde{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{17{,}5 + 18{,}8}{9} = 4{,}03.$$

Середню помилку різниці двох середніх при неоднаковій кількості спостережень обчислюють за формулою

$$\overline{m}_{1-2} = \sqrt{\sigma_0^2 \frac{n_1 + n_2}{n_1 \cdot n_2}} = \sqrt{\frac{4{,}03 \cdot 11}{30}} = 1{,}2.$$

Фактичне значення критерію

$$t_\phi = \frac{x_1 - x_2}{\overline{m}_{1-2}} = \frac{1{,}7}{1{,}2} = 1{,}4167.$$

Табличне значення t_T при рівні значущості 0,05 і 9 ступенях свободи варіації дорівнює 2,2622. Оскільки $t_\phi < t_T$, то нульова гіпотеза не відхиляється і розбіжності між середніми не можна визнати істотними.

Оцінка вірогідності середньої різниці. Якщо вибірки сформовані як залежні, то в цьому разі проводять статистичну оцінку середньої різниці. Послідовність розрахунків розглянемо на прикладі досліду, в якому вивчалась урожайність двох сортів озимої пшениці — Донецької 46 і Харківської 11 (табл.32). Обидва сорти утворюють пари, і врожайність їх можна порівняти тільки попарно в межах кожної повторності.

Щоб оцінити результати порівняння сортів озимої пшениці, висунемо гіпотезу про те, що одержана середня різниця $\overline{d}$ випадкова, і висновок гіпотези зробимо з рівнем значущості 0,05.

Таблиця 32 *Урожайність озимої пшениці, ц/га*

Повторність	Донецька 46	Харківська 11	Різниця урожайності	Квадрат різниці
	x_1	x_2	d	d^2
1	51,8	55,3	3,5	12,25
2	52,4	57,1	4,7	22,09
3	47,8	51,8	4,0	16,00
4	47,2	50,7	3,0	9,00
Разом	199,2	214,4	15,2	59,34
Середні	49,8	53,6	3,8	—

Обчислимо середню різницю:

$$\overline{d} = \frac{\sum d}{n} = \frac{15{,}2}{4} = 3{,}8 \text{ ц/га}.$$

Знайдемо дисперсію окремих різниць d від середньої різниці $\overline{d}$:

$$\sigma_d^2 = \frac{\sum d^2 - n\overline{d}^2}{n-1} = \frac{59{,}34 - 4 \cdot 3{,}8^2}{4-1} = 0{,}53.$$

Середня помилка середньої різниці

$$m_d = \sqrt{\frac{\sigma_d^2}{n}} = \sqrt{\frac{0{,}53}{4}} = 0{,}36 \text{ ц/га}.$$

Порівняємо середню різницю із середньою помилкою, тобто визначимо фактичне значення *t*-критерію:

$$t_ф = \frac{\bar{d}}{\bar{m}_d} = \frac{3,8}{0,30} = 10,56.$$

Табличне значення *t*-критерію (дод. 2) при рівні значущості 0,05 і 3 ступенях свободи варіації дорівнює 3,18.

Оскільки фактичне значення критерію Стьюдента більше за табличне, то нульова гіпотеза відхиляється.

Знаючи середню помилку різниці і табличне значення критерію, обчислимо граничну помилку середньої різниці:

$$\Delta = t\bar{m}_d = 3,18 \cdot 0,36 = 1,1 \text{ ц/га}.$$

Гранична помилка показує, що внаслідок випадкових коливань середня різниця урожайності порівнюваних сортів може досягти 1,1 ц/га. Фактична ж середня різниця становить 3,8 ц/га. Отже, можна зробити висновок, що розмір фактичної середньої різниці не узгоджується з нульовою гіпотезою. Тому приймаємо альтернативну гіпотезу: сорт озимої пшениці Харківська 11 має більш високу врожайність порівняно із сортом Донецька 46.

4.4 Перевірка статистичних гіпотез відносно розподілів частот

Для того щоб перевірити відповідність двох порівнюваних рядів розподілу частот (емпіричного і теоретичного), використовують критерій χ^2. Величину χ^2 обчислюють за формулою

$$\chi^2 = \sum \frac{(f - f')^2}{f'},$$

де f — фактичні (емпіричні) частоти розподілу; f' — очікувані (теоретично обчислені) частоти розподілу.

Величина χ^2 є середньою зваженою квадратів відхилень фактичних і теоретичних частот. При цьому вагами є величини, обернені теоретичним частотам. Чим більша розбіжність між фактичними і теоретичними частотами, тим більша величина χ^2. Вона може мати значення від 0 до ∞. Якщо фактичні і теоретичні частоти однакові $(f - f' = 0)$, значення χ^2 дорівнюватиме нулю.

Крива розподілу χ^2 змінюється залежно від кількості ступенів свободи варіації. При великій кількості спостережень розподіл χ^2 набуває форми, близької до нормального.

Відхилення фактичних частот від теоретичних може бути зумовлене випадковими причинами або відображати існуючі розбіжності між емпіричним і теоретичним розподілом. Щоб визначити істотне чи випадкове відхилення, ожержані у досліді значення χ^2 порівнюють з табличними (дод. 6). У цих таблицях в основній графі показують кількість ступенів свободи, а у загальному рядку — рівень значущості. На перетині їх маємо значення χ^2, яке відповідає цій кількості ступенів свободи і рівню ймовірності. Наприклад, при 10 ступенях свободи і рівні значущості 0,05 критерій χ^2 становитиме 18,31. Це означає, що при 10 ступенях свободи значення

χ^2 унаслідок дії випадкових причин може дорівнювати 18,31 або бути більшим у 5 випадках із 100.

Якщо фактичне значення обчисленого за даними вибірки критерію χ^2 дорівнює табличному або менше за нього (при відповідній кількості ступенів свободи і рівні значущості), то це означає, що розбіжності між фактичними і теоретичними частотами випадкові. Якщо фактичне значення χ^2 більше за табличне, то це означає, що розбіжності між емпіричними і теоретичними частотами зумовлені не випадковими, а істотними причинами.

За допомогою критерію χ^2 оцінюють узгодження (відповідність) між фактичним і теоретичним розподілом частот, незалежність розподілу одиниць сукупності за градаціями досліджуваної ознаки, а також однорідність розподілу.

При перевірці гіпотези про відповідність емпіричного розподілу теоретичному за допомогою критерію χ^2 потрібно мати не менш як 50 спостережень, а груп з теоретичною частотою менше 5 одиниць повинно бути не більш як 20 % обсягу вибірки.

Кількість ступенів свободи для визначення табличного значення χ^2 при дослідженні відповідності фактичного розподілу теоретичному дорівнює кількості груп без трьох. Це зумовлене тим, що обчислення теоретичних частот пов'язане з трьома умовами, які визначають нормальний розподіл: обсягом вибірки N, середнім значенням ознаки $\bar{x}$ і дисперсією σ^2.

Досліджуючи незалежність розподілу, ступені свободи варіації для визначення табличного значення χ^2 встановлюють залежно від кількості груп і підгруп, на які розподіляється сукупність. Якщо a — кількість груп, b — кількість підгруп, то $\nu = (a-1)(b-1)$.

Розглянемо послідовність розрахунків при оцінці узгодження між фактичним і теоретичним розподілом частот. Наприклад, є дані про розподіл 200 корів племінного господарства за живою масою (табл. 33). Потрібно виявити вплив живої маси корів на їх продуктивність за допомогою кореляційного аналізу. Необхідною умовою проведення кореляційного аналізу є нормальний розподіл сукупності. У

зв'язку з цим слід перевірити відповідність емпіричного розподілу нормальному за допомогою критерію узгодження χ^2.

Сформулюємо гіпотези: нульову — емпіричний розподіл відповідає нормальному; альтернативну — емпіричний розподіл не відповідає нормальному.

Таблиця 33 *Обчислення теоретичних частот і критерію* χ^2

Вихідні дані			Розрахункові дані					
Жива маса корови, кг	середина інтервалу x	кількість корів, f	$t=\frac{x-\bar{x}}{\sigma}$	$\Phi(t)$	f'	$f-f'$	$(f-f')^2$	$\frac{^3(f-f')}{f'}$
400 – 420	410	5	–2,34	0,0258	4	1	1	0,25
420 – 440	430	19	–1,58	0,1145	17	2	4	0,24
440 – 460	450	38	–0,83	0,2827	43	–5	25	0,58
460 – 480	470	63	–0,08	0,3977	60	3	9	0,15
480 – 500	490	44	0,68	0,3166	48	–4	16	0,33
500 – 520	510	26	1,43	0,1435	23	3	9	0,39
520 – 540	530	5	2,19	0,0363	5	0	0	0
Разом	—	200	—	—	200	—	—	1,94

Для розв'язання задачі потрібно за даними емпіричного розподілу визначити теоретичні частоти, які відповідають кривій нормального розподілу. Відомо, що крива нормального розподілу характеризується двома параметрами: середньою і середнім квадратичним відхиленням.

У нашому прикладі середня жива маса корів становить:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{94400}{200} = 472 \text{ грн.}$$

Обчислимо середнє квадратичне відхилення з урахуванням поправки Бесселя за формулою

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2 f}{\sum f - 1}} = \sqrt{\frac{140000}{199}} = 26{,}5.$$

Знаючи середню і середнє квадратичне відхилення живої маси, визначимо нормовані відхилення:

$$t = \frac{x - \bar{x}}{\sigma},$$

які запишемо у четверту графу табл. 33.

Користуючись додатку 7, знаходимо $\Phi(t)$ – ймовірність повторення даного значення ознаки х у нормально розподіленій сукупності. Наприклад, значенню $t = 2{,}34$ відповідає ймовірність повторення 0,0258 (258 сприятливих наслідків із 10000).

Обчислимо теоретичні частоти за формулою

$$f' = \Phi(t)\frac{N \cdot i}{\sigma},$$

де N — чисельність сукупності; i — величина інтервалу.

Постійний множник $\frac{N \cdot i}{\sigma}$ і у нашому прикладі дорівнюватиме $\frac{200 \cdot 20}{26{,}5} = 150{,}9$.

У нашій групі теоретичні частоти становлять 0,0258 · 150,9 = 4.

Підсумовуючи дані останьої графи, матимемо фактичне значення критерію узгодження $\chi^2 = 1{,}94$.

Кількість ступенів свободи варіації визначають як кількість груп у ряді $n = 7$ мінус кількість показників емпіричного ряду, використаних при обчисленні теоретичних частот. У нашому прикладі таких показників три: $\bar{x}, \sigma$ і N, звідки $v = n - 3 = 7 - 3 = 4$. При чотирьох ступенях свободи і рівні значущості 0,05 табличне значення χ^2 дорівнює 9,49. Оскільки $\chi^2_\phi < \chi^2_т$, то нульова гіпотеза приймається. Робимо висновок, що емпіричному розподілу корів за живою масою відповідає теоретичний, а розбіжності між ними зумовлені випадковими причинами.

Таблиця 34 *Ураженість озимої пшениці кореневою гниллю залежно від виду пару*

Вид пару	Ураженість		Сума	Процент слабко уражених полів
	слабка	сильна		
Чистий	31	10	41	75,6
Зайнятий	37	37	74	50,0
Сума	68	47	115	59,1

Розглянемо тепер використання χ^2 як критерію незалежності розподілу одиниць сукупності за градаціями досліджуваної ознаки. Припустімо, що після обстеження 115 полів озимої пшениці на ураженість кореневою гниллю одержано такі дані.Потрібно оцінити істотність змін у співвідношенні слабкої і сильної ураженості полів кореневою гниллю. Рівень значущості 0,01.

Висунемо гіпотезу про те, що вид пару не змінює співвідношення слабкої і сильної ураженості полів. Цій гіпотезі відповідатиме

рівномірний розподіл частот по групах відповідно до підсумкового розподілу, тобто 59,1 % слабко уражених полів і 40,9 % сильно уражених. Для того, щоб обчислити теоретичні (очікувані) частоти, загальну кількість полів у кожній групі множать на очікувану частоту слабко уражених (59,1 %). Кількість сильно вражених полів визначають як різницю між загальною кількістю полів у групі і кількістю слабко уражених (табл. 35).

Таблиця 35 *Очікувана кількість полів, слабко і сильно уражених кореневою гниллю*

Вид пару	Ураженість		Сума	Процент слабкоуражених полів
	слабка	сильна		
Чистий	24	17	41	59,1
Зайнятий	44	30	74	59,1
Сума	68	47	115	59,1

В нижче наведеній таблиці визначимо різниці між фактичними й очікуваними частотами (табл. 36).

Таблиця 36 *Різниця між фактичними і очікуваними частотами полів*

Вид пару	Ураженість		Разом
	слабка	сильна	
Чистий	7	–7	0
Зайнятий	–7	7	0
Разом	0	0	0

Фактичне значення критерію

$$\chi^2 = \frac{(f - f')^2}{f'} = \frac{7^2}{24} + \frac{(-7)^2}{17} + \frac{(-7)^2}{44} + \frac{7^2}{30} = 7{,}66.$$

Кількість ступенів свободи варіації:

$$v = (a - 1)(b - 1) = (2 - 1)(2 - 1) = 1.$$

Табличне значення χ^2 при одному ступені свободи варіації і рівні значущості 0,01 дорівнює 6,63. Оскільки фактичне значення критерію χ^2 перевищує табличне, то висунута нульова гіпотеза про незалежність ураженості полів кореневою гниллю від пару не узгоджується з фактичними спостереженнями. При розміщенні озимої пшениці по зайнятому пару ураженість полів кореневою гниллю істотно збільшується.

Статистичні гіпотези про однорідність двох сукупностей з використанням χ^2 як критерію однорідності перевіримо на такому прикладі (табл. 37). Перевіримо гіпотезу про однорідність двох емпіричних розподілів.

Спочатку сформулюємо гіпотези: нульову — розподіли однорідні, тобто взяті з однієї генеральної сукупності; альтернативну — розподіли неоднорідні, тобто взяті з різних генеральних сукупностей.

Приймемо рівень значущості 0,05.

Для перевірки нульової гіпотези слід обчислити фактичне значення критерію χ^2 і порівняти його з табличним значенням $\chi^2_{0,05}$.

Послідовність розрахунків фактичного значення χ^2 показана в останніх шести графах табл. 37. Щоб визначити фактичне значення χ^2, слід підсумок графа 7 помножити на $n_1 n_2$:

Таблиця 37 *Розподіл д за стажем роботи у державних господарствах двох районів і розрахунок критерію однорідності*

Група доярок за стажем роботи	Кількість доярок			Частка групи доярок у загальній їх кількості		Різниця	Квадрат різниці	Вага	Зважений квадрат різниці
	n_1	n_2	n_1+n_2	$d_1=\frac{n_1}{\sum n_1}$	$d_2=\frac{n_2}{\sum n_2}$	$L=d_1-d_2$	l^2	$\frac{1}{n_1+n_2}$	$l^2\cdot\frac{1}{n_1+n_2}$
0 — 10	24	15	39	0,20	0,15	0,05	0,0025	0,0256	0,000064
10 — 20	48	30	78	0,40	0,30	0,10	0,0100	0,0128	0,000128
20 — 30	42	45	87	0,35	0,45	–0,10	0,0100	0,0115	0,000115
30 — 40	6	10	16	0,05	0,10	–0,05	0,0025	0,0625	0,000156
Разом	120	100	220	1,00	1,00	—	—	—	0,000465

$$\lambda_\phi^2 = n_1 n_2 \sum l^2 \frac{1}{n_1+n_2} = 120 \cdot 100 \cdot 0{,}000463 = 5{,}56.$$

Кількість ступенів свободи варіації:

$$K = a - 1 = 4 - 1 = 3.$$

де а — кількість груп (інтервалів).

Табличне значення χ^2 при 3 ступенях свободи варіації і рівні значущості 0,05 (дод. 6) дорівнює 7,82.

Оскільки $\chi^2_\phi < \chi^2_{0,05}$ (5,56 < 7,72), слід погодитися з нульовою гіпотезою. Досліджувані групи доярок можна вважати однорідними за стажем роботи, оскільки вони взяті з однієї генеральної сукупності.

4.5 Перевірка статистичних гіпотез за допомогою непараметричних критеріїв

У статистичній практиці з непараметричних критеріїв найчастіше використовують критерії Вілкоксона та Колмогорова. Критерій Вілкоксона застосовують для оцінки істотності відмінностей вибірок з попарно пов'язаними варіантами, критерій Колмогорова — для оцінки вірогідності розбіжностей між фактичним і теоретичним розподілами частот.

Перевірку статистичних гіпотез за допомогою критерію Вілкоксона розглянемо на прикладі (табл. 38).

Таблиця 38 *Урожайність картоплі в господарствах району до і після впровадження орендного підряду*

Номер господарства	Урожайність картоплі у підрозділах, ц/га			
	які впровадили орендний підряд		які не впровадили орендний підряд	
	у базисному періоді	у звітному періоді	у базисному періоді	у звітному періоді
1	207	235	214	220
2	189	211	185	190
3	214	227	210	226
4	224	230	219	230
5	190	217	192	207
6	185	197	197	215
7	214	238	205	210
8	175	208	180	195
9	181	216	176	186
10	196	224	200	214

Таблиця 39 *Різниці в урожайності картоплі та їх порядкові ранги*

Номер господарства	1	2	3	4	5	6	7	8	9	10
Різниця в урожайності	+22	+17	-3	-5	+18	-6	+19	+18	+25	+14
Ранг різниці	9	5	-1	-2	6	-3	8	7	10	4

Наведені в табл 39 дані стосуються періодів, коли була тільки одна форма організації праці (базисний період) і коли у кожному господарстві паралельно із звичайною впровадили орендну форму організації праці (звітний період). Потрібно оцінити істотність змін в урожайності картоплі у зв'язку з упровадженням орендного підряду. Рівень значущості 0,05.

Для перевірки гіпотези обчислимо різниці урожайності у кожному господарстві за формулою:

$$d = \left(J_1' - J_0'\right) - \left(J_1 - J_0\right)$$

де J'_1 і J'_0 — урожайнсть картоплі у підрозділах, які впровадили орендний підряд у звітному та базисному періодах; J_1 і J_0 — урожайність картоплі у підрозділах із звичайною формою організації праці у звітному та базисному періодах.

Обчислені різниці урожайності картоплі та їх порядкові ранги з відповідним знаком запишемо у табл. 39.

Визначимо суми рангів з додатним і від'ємним знаками:

$$+T = 49; -T = 6.$$

Меншу суму рангів (без урахування знака) порівнюють з табличним значенням критерію Вілкоксона (дод. 8) при вибраному рівні значущості. У нашому прикладі критичне значення критерію Вілкоксона при 10 попарно пов'язаних спостереженнях і рівні значущості 0,05 дорівнює 8.

Оскільки фактична сума від'ємних рангів менша за табличне значення ($6 < 8$), то нульова гіпотеза відхиляється на відповідному рівні значущості. З імовірністю помилитися у 5 випадках із 100 можна стверджувати, що розбіжності в урожайності картоплі є наслідком впровадження орендного підряду.

Для первірки статистичних гіпотез про вірогідність розбіжностей між фактичним і теоретичним розподілами, а також відмінностей між якими-небудь двома фактичними розподілами частот певної ознаки використовують λ-критерій Колмогорова. Цей критерій грунтується на порівнянні рядів інтегральних (нагромаджених) частот сукупностей. Якщо порівнювані вибірки належать до тієї самої генеральної сукупності, то різниця між частотами буде мінімальною і буде зумовлена випадковими коливаннями дослідних даних. Велике максимальне значення різниці між частотами є підставою для відхилення нульової гіпотези та визнання відмінностей між вибірками істотними.

Перевірку гіпотези про відповідність емпіричного розподілу нормальному розглянемо за допомогою критерію Колмогорова на такому прикладі (табл. 40).

Сформулюємо гіпотези: нульову — емпіричний розподіл відповідає нормальному; альтернативну — емпіричний розподіл не відповідає нормальному. Рівень значущості = 0,05.

Обчислимо фактичне значення λ-критерію Колмогорова за формулою

$$\lambda = \frac{d}{\sqrt{n}} = \frac{4}{\sqrt{200}} = 0{,}28$$

де d — найбільша різниця (без урахування знака) між нагромадженими частотами в емпіричному і теоретичному розподілах; n — чисельність вибірки.

Послідовність розрахунку λ-критерію показана в останніх трьох графах табл. 40.

Таблиця 40 *Емпіричний і теоретичний розподіл 200 поросят при відлученні за живою масою*

Жива маса поросят, кг	Частоти		Нагромаджені частоти		Різниця між нагромадженими частотами d
	емпіричні	теоретичні	емпіричні	теоретичні	
12	7	5	7	5	2
13	10	11	17	16	1
14	24	23	41	39	2
15	30	36	71	75	4
16	45	43	116	118	2
17	36	38	152	156	2
18	32	25	184	181	3
19	11	13	195	194	1
20	5	6	200	200	0
Разом	200	200	—	—	—

Фактичне значення λ-критерію = 0,28 значно менше за критичний $\lambda_{0,05} = 1{,}36$. Отже, слід погодитися з нульовою гіпотезою, яка передбачає відповідність емпіричного розподілу нормальному. Вибірковий розподіл поросят за живою масою узгоджується з нормальним.

4.6 Перевірка статистичних гіпотез відносно частки ознаки

Частки двох вибіркових розподілів якісних ознак порівнюють за допомогою t-критерію нормального розподілу при великих вибірках і крітерію t-Стьюдента при малих вибірках, тобто через відношення різниці між перевірюваними частками до помилки цієї різниці.

Помилку різниці двох вибіркових часток визначають за формулою

$$m = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},$$

де p_1 і p_2 — частка і кількість спостережень одного розподілу, p_2 і n_2 — частка і кількість спостережень другого розподілу.

Перевірку гіпотези відносно частки ознаки розглянемо на такому прикладі. У двох районах проведено вибіркове опитування автомобілістів з метою визначення їх ставлення до якості технічного обслуговування автомобілів підприємствами автосервісу. У першому районі з 300 опитаних автомобілістів позитивно оцінили якість технічного обслуговування 168 чоловік, негативно — 132. У другому районі з 280 опитаних позитивні відповіді дали 126, негативні — 154 автомобілісти. Потрібно перевірити гіпотезу про істотність відмінностей стосовно частки позитивних відповідей у генеральній сукупності.

Сформулюємо нульову й альтернативну гіпотези:

$$H_0: p_1 = p_2;\ \text{На: } p_1 \neq p_2.$$

Рівень значущості — 0,05.

Оскільки маємо справу з великою вибіркою, то найбільш потужним критерієм перевірки цієї гіпотези є t-критерій нормального розподілу.

Визначимо частку позитивних (p) і негативних ($1 - p$) відповідей в обох вибірках:

$$p_1 = \frac{168}{300} = 0{,}56;\ 1 - p_1 = 0{,}44;$$
$$p_2 = \frac{126}{280} = 0{,}45;\ 1 - p_2 = 0{,}55.$$

Фактичне значення t-критерію обчислемо за формулою

$$t_\phi = \frac{(p_1 - p_2)}{\sqrt{\frac{p_1(1-p)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{(0{,}56 - 0{,}45)}{\sqrt{\frac{0{,}56 \cdot 0{,}44}{300} + \frac{0{,}45 \cdot 0{,}55}{280}}} = 2{,}65.$$

Табличне значення t-критерію нормального розподілу (дод. 1) при рівні значущості 0,05 дорівнює 1,96.

Оскільки $t_\phi > t_{0,05}$ (2,65 > 1,96) нульова гіпотеза про рівні частки позитивних відповідей у генеральній сукупності відхиляється. Розбіжності часток не є випадковими.

Контрольні питання

1. *Що таке статистичні гіпотези та які існують їх види?*
2. *Які помилки виникають при перевірці статистичних гіпотез?*
3. *Що таке статистичні критерії та які існують їх види?*
4. *Що таке критична галузь, галузь допустимих значень і критичні точки?*
5. *Яка принципова схема перевірки статистичних гіпотез?*
6. *Як здійснюється перевірка гіпотез про істотність розбіжностей між дисперсіями?*
7. *Як здійснюється оцінка вірогідності різниці двох середніх?*
8. *Як здійснюється оцінка вірогідності середньої різниці?*
9. *Як перевіряються гіпотези відносно розподілу частот?*
10. *Як перевіряються гіпотези відносно частки ознаки?*

Розділ 5
ДИСПЕРСІЙНИЙ АНАЛІЗ

5.1 Основи та принципова схема дисперсійного аналізу

Для статистичної оцінки взаємозв'язків між явищами та їх істотності при невеликій кількості спостережень застосовують дисперсійний аналіз.

Дисперсійний аналіз — це метод оцінки впливу одного чи кількох факторів, що одночасно діють на певну результативну ознаку. Його застосовують під час статистичної обробки даних, одержаних унаслідок експерименту або спостереження, для виявлення впливу окремих факторів та їх взаємодії на рівень показників ефективності сільськогосподарського виробництва. Такий аналіз дає змогу кількісно характеризувати вплив на результативну ознаку різних факторів, навіть тих, які не виражаються числом, а належать до атрибутивних ознак. При цьому атрибутивні ознаки можна вивчати разом з кількісними.

Розробка методу дисперсійного аналізу пов'язана з ім'ям англійського статистика Р.А.Фішера, який уперше застосував його у 1925 р. при обробці результатів експериментальних досліджень.

Дисперсійний метод аналізу найчастіше використовують при розробці результатів багатоваріантних дослідів для загальної оцінки вірогідності розбіжностей у групових середніх, групуючи дані за однією чи кількома факторними ознаками, а також для визначення вірогідності взаємодії двох, трьох або більшої кількості факторів.

При дисперсійному аналізі досліджувані об'єкти зводять у дисперсійні комплекси. Вони становлять прості або комбіновані групування, в яких для кожного фактора виділяють щонайменше дві групи (градації). Залежно від кількості досліджуваних факторів дисперсійні комплекси бувають одно-, дво- і трифакторними.

На варіацію досліджуваних ознак впливають різноманітні фактори, які поділяють на систематичні та випадкові. У зв'язку із цим розрізняють варіацію систематичну та випадкову. **Систематична варіація** — це частина загальної варіації результативної

ознаки, зумовлена систематичною дією факторних ознак (наприклад, різною родючістю грунтів, різними дозами внесених добрив, різним рівнем годівлі тварин тощо). **Випадкова варіація** — це частина загальної варіації результативної ознаки, зумовлена дією випадкових факторів. Випадкову варіацію часто називають **залишковою**, оскільки вона відображує варіацію результативної ознаки, зумовлену іншими причинами, не врахованими в обсязі систематичної варіації.

Суть дисперсійного аналізу — в розподілі загальної варіації досліджуваної ознаки на систематичну та випадкову (залишкову) і в порівнянні систематичної варіації з випадковою.

Обсяг варіації результативної ознаки в дисперсійному комплексі визначають сумами квадратів відхилень: загальною, міжгруповою (систематичною, факторною) і залишковою (випадковою). Діленням відповідної суми варіації на кількість ступенів свободи встановлюють загальну, міжгрупову та залишкову дисперсії.

Кількість ступенів свободи варіації характеризує кількість елементів сукупності, що вільно варіюють. Так, якщо на основі вибірки $x_1, x_2, \ldots, x_n$ розрахована їх середня арифметична $\bar{x}$, то сукупність відхилень $x_i - \bar{x}(i = 1,\ldots,n)$ має $n-1$ ступенів свободи. Кількість ступенів свободи завжди буде меншою на одиницю від кількості варіюючих величин, серед яких обчислена середня. Якщо сума 10 чисел дорівнює 50, то 9 з них можуть бути будь-які довільні, а десяте число буде фіксованим (як різниця між 50 і сумою 9 довільних чисел), тобто воно втрачає свою свободу.

Залишкову дисперсію, зумовлену дією випадкових факторів, при дисперсійному аналізі використовують як міру помилки спостереження. При цьому міжгрупову дисперсію порівнюють із залишковою. Якщо міжгрупова варіація істотно більша за залишкову, то це означає, що вона не випадкова, а зумовлена впливом фактора, що вивчається. Нормативом, з яким проводять порівняння для оцінки вірогідності відношення систематичної дисперсії до випадкової, є математичний критерій F, розроблений Р.Фішером.

Критерій F — це співвідношення двох дисперсій:

$$F = \frac{S_1^2}{S_2^2},$$

де S_1^2 і S_2^2 — дисперсії (середні квадрати відхилень), обчислені за вибірковими даними з урахуванням кількості ступенів свободи варіації.

Якщо вибірки з дисперсіями S_1^2 і S_2^2 здійснено з тієї самої генеральної сукупності, де варіація величин мала випадковий характер, то і розбіжність у дисперсіях S_1^2 і S_2^2 також випадкова.

Р.Фішер розробив спеціальні математичні таблиці, в яких наводиться величина F для вибірок, зроблених з тієї самої сукупності, де причини зміни величин мають тільки випадковий характер. Теоретичне значення F показує можливу випадкову величину співвідношення двох дисперсій при відповідному рівні ймовірності і ступенів свободи варіації.

У дод. 5 наведено теоретичне значення F при ймовірності 0,95 (5% рівень розподілу). Рівень ймовірності 0,95 показує, що тільки в 5 випадках із 100 значення F може бути більше у випадковій вибірці. Підвищення рівня ймовірності (наприклад, до 0,99 чи 0,999) призводить до збільшення F.

Таблиця значень F побудована так: у графах зазначено ступені свободи варіації для великої дисперсії, а в рядках — ступені свободи варіації для малої дисперсії. Величина F міститься на перетині графи та рядка, які відповідають ступеням свободи варіації порівнюваних дисперсій. Якщо для великої дисперсії v (ступені сводоби варіації) дорівнюють 7, а для малої — 9, то табличне значення F при ймовірності 0,95 становить 3,29. Це означає, що дисперсія однієї вибірки може перевищувати дисперсію другої вибірки при ймовірності 0,95 у 3,29 раза.

Табличне значення F використовують як критерій для оцінки фактичних співвідношень дисперсій, обчислених за вибірковими даними. Якщо фактичне значення F більше за табличне, то це свідчить про достатню вірогідність впливу факторної ознаки на результативну. У цьому разі розбіжності у варіації величин пов'язані не тільки з випадковим коливанням їх рівня і різниця між порівнюваними дисперсіями вважається ймовірною. І навпаки, якщо фактичне значення F менше за теоретичне, то розбіжності між порівнюваними дисперсіями лежать у межах можливих випадкових коливань.

Дисперсійний аналіз здійснюють за принциповою схемою:

1) визначення джерел варіювання досліджуваної ознаки і добір моделі дисперсійного аналізу;
2) обчислення обсягів варіації за джерелами утворення;
3) визначення кількості ступенів свободи варіації;
4) розрахунок дисперсії та показників співвідношень між ними;
5) аналіз співвідношень між дисперсіями і загальні висновки. Цю схему використовують як при простих моделях, коли групувальна ознака одна, так і при складни, коли групувальних ознак дві або більше.

При застосуванні дисперсійного аналізу слід дотримуватися таких умов. Досліджувана сукупність повинна мати нормальний або близький до нього розподіл. Нерідко при досліджуванні соціально-економічних явищ розподіли одиниць сукупності мають асиметричний характер. Це призводить до того, що оцінки, одержані на основі F-критерію, фактично мають менший рівень вірогідності, ніж свідчать розрахунки. Крім того, окремі спостереження в групах і підгрупах мають бути незалежними. При використанні узагальнених показників, що характеризують складні соціально-економічні явища, ця умова, як правило, порушується, що призводить до можливих помилок у висновках.

Застосування певної моделі дисперсійного аналізу залежить як від кількості досліджуваних факторів (одно-, дво- і трифакторні моделі), так і від способу формування груп (випадкове та невипадкове формування).

5.2 Дисперсійний аналіз при групуванні даних за однією ознакою

Послідовність і техніку проведення дисперсійного аналізу при групуванні даних за однією ознакою розглянемо на такому прикладі (табл. 42).

Таблиця 42 *Вплив зростаючих доз фосфорних добрив на врожайність цукрових буряків, ц/га*

Варіант досліду	Урожайність на ділянках				Середня врожайність
	1	2	3	4	
Контроль	365	379	378	322	361
P_{20}	368	418	398	363	386
P_{40}	383	421	400	381	396
P_{60}	420	431	411	407	417
P_{80}	399	433	411	413	414
P_{100}	430	421	410	402	416

Дані таблиці показують, що із збільшенням доз фосфорних добрив підвищується середня урожайність цукрових буряків за варіантами досліду. Отже, варіація середніх по групах зумовлюється різними дозами фосфорних добрив. Одночасно спостерігається коливання врожайності цукрових буряків на ділянках у кожному варіанті досліду. Так, при внесенні 20 кг фосфорних добрив урожайність на ділянках коливається від 263 до 418 ц, при внесенні 40 кг — від 381 до 421 ц. Звідси на рівень урожайності, крім унесення фосфорних добрив, впливають й інші фактори, що зумовлює внутрішньогрупову варіацію.

Загальний обсяг варіації, який визначають сумою квадратів відхилень індивідуальних значень урожайності на ділянках від загальної середньої урожайності всього досліду, дорівнює сумі міжгрупової та внутрішньогрупової варіації:

$$W_{заг.} = W_{м.гр.} + W_{вн.гр.}$$

Порядок обчислення сум квадратів відхилення детально розглянуто в розділі 2. Щоб спростити розрахунки при визначенні квадратів та їх сум, усі значення врожайності зменшимо на довільне число 400, тобто скористаємося способом розрахунку від умовного нуля (табл. 43).

Таблиця 43 *Відхилення від умовного нуля (y =x – a, a = 400)*

Варіант досліду	Урожайність на ділянках				$\sum y_j$
	1	2	3	4	
Контроль	–35	–21	–22	–78	–156
P_{20}	–32	18	–2	–27	–53
P_{40}	–17	21	0	–19	–15
P_{60}	20	31	11	7	69
P_{80}	–1	33	11	13	56
P_{100}	30	21	10	5	63
$\sum y_i$	–35	103	8	–112	$\sum y_{ij} = 36$

Обчислені відхилення від умовного нуля (y) і суми відхилень піднесемо до квадрата (табл. 44).

На підставі даних табл. 44 визначаємо суми квадратів відхилень. При цьому загальну кількість спостережень позначимо

через $N = 24$, кількість варіантів досліду — через $m \neq 6$, кількість ділянок — через $n = 4$.

Таблиця 44 *Розрахунок квадратів відхилень*

Варіант досліду	Урожайність на ділянках, ц/га				Сума квадратів $\sum y_j^2$	Квадрат суми $(\sum y_j)^2$
	1	2	3	4		
Контроль	1225	441	484	6084	8234	24336
P_{20}	1024	324	4	1369	2721	2809
P_{40}	289	441	0	361	1091	225
P_{60}	400	961	121	49	1531	4761
P_{80}	1	1089	121	169	1380	3136
P_{100}	900	441	100	4	1445	3969
Сума квадратів $\sum y_i^2$	3839	3697	830	8036	$\sum y_{ij}^2 = 16402$	39236
Квадрат суми $(\sum y_i)^2$	1225	10609	64	12544	24442	—

Загальна сума квадратів відхилень

$$W_{заг} = \sum y_{ij}^2 - \frac{(\sum y_{ij})^2}{N} = 16402 - \frac{36}{24} = 16348.$$

Сума квадратів відхилень середньої врожайності окремих груп (варіантів досліду) від загальної середньої

$$W_{доб.} = \frac{\sum\left(\sum y_j\right)^2}{n} - \frac{\left(\sum y_{ij}\right)^2}{N} = \frac{39236}{4} - \frac{36}{24} = 9755.$$

Сума квадратів відхилень середньої врожайності на окремих ділянках від загальної середньої

$$W_{діл} = \frac{\sum\left(\sum y_i\right)^2}{m} - \frac{\left(\sum y_{ij}\right)^2}{N} = \frac{24442}{6} - \frac{36}{24} = 4020.$$

Суму квадратів відхилень залишкової варіації обчислимо як різницю між загальною варіацією та варіацією, зумовленою добривами та рівнем урожайності на окремих ділянках:

$$W_{зал} = W_{заг} - W_{доб} - W_{діл} = 16348 - 9755 - 4020 = 2573.$$

Згідно з принциповою схемою дисперсійного аналізу визначаємо ступені свободи варіації (v) для кожної суми квадратів відхилень. Ступені вільності характеризують кількість елементів сукупності, що вільно варіюють. Якщо серед варіюючих величин обчислено середню, то незалежних відхилень від середньої буде на одиницю менше, тобто кількість ступенів свободи завжди менша на одиницю від кількості варіюючих величин, серед яких обчислено середню. У нашому прикладі

$$V_{заг} = N - 1 = 24 - 1 = 23;$$
$$V_{доб} = m - 1 = 6 - 1 = 5;$$
$$V_{діл} = n - 1 = 4 - 1 = 3.$$

Кількість ступенів свободи варіації для залишкової суми квадратів відхилень визначають як різницю між загальною кількістю ступенів свободи і ступенів свободи варіації, зумовленої добривами та рівнем урожайності на окремих ділянках:

$$V_{зал} = (N - 1) - (m - 1) - (n - 1) = 23 - 5 - 3 = 15.$$

Знаючи суми квадратів відхилень і ступені вільності, можна обчислити різні види дисперсії (σ^2). Для цього суми квадратів відхилень потрібно розділити на кількість ступенів свободи варіації (табл.45).

Таблиця 45 *Аналіз дисперсій*

Джерела варіації	Сума квадрів відхилень	Кількість ступенів свободи	Дисперсія	Відношення дисперсій	
				F_ϕ	F_T при ймовірності 0,95
Добрива	9755	5	1951	11,38	2,90
Ділянки	4020	3	1340	7,81	3,29
Залишкова	2573	15	171,5	1	—
Загальна	16348	23	—	—	—

У табл. 45 фактичне значення F для добрив (11,38) обчислене як відношення дисперсії добрив до залишкової дисперсії, а для ділянок (7,81) — як відношення дисперсії ділянок до залишкової дисперсії.

Табличне значення F для оцінки відношення дисперсії добрив до залишкової дисперсії знаходимо у дод. 5 на перетині графи 5 і рядка 15 ($F_{0,95} = 2,90$), а для відношення дисперсії ділянок до залишкової — на перетині графи 3 і рядка 15 ($F_{0,95} = 3,29$).

Фактичне відношення між порівнюваними дисперсіями значно перевищує табличне. Це дає змогу зробити висновок, що різниця між середніми рівнями врожайності як по добривах, так і по ділянках не пов'язана з випадковим варіюванням, а є істотною, вірогідною.

При дисперсійному аналізі залишкову дисперсію називають **дисперсією помилки**. Вона характеризує варіацію результативної ознаки під впливом випадкових факторів, не врахованих у досліді. Чим менше буде залишкова дисперсія, тим менше в досліді випадкових помилок і тим він точніший.

Наведену вище схему дисперсійного аналізу використовують тоді, коли вибірка сформована як залежна, тобто розподіл ділянок у кожному варіанті не випадковий. При цьому варіювання дослідних даних зумовлене не тільки особливостями варіантів, а й різною родючістю грунту окремих ділянок, особливостями досліджуваних рослин тощо.

Якщо вибірка сформована як незалежна, тобто розподіл ділянок у кожному варіанті випадковий, то загальну суму квадратів відхилень поділяють на два компоненти:

1) суму квадратів відхилень середньої врожайності у кожному варіанті досліду від загальної середньої, тобто факторну варіацію. Ця сума квадратів відхилень характеризує варіацію, зумовлену дією фактора, що вивчається;
2) суму квадратів відхилень індивідуальних значень урожайності у кожному варіанті досліду від групової середньої, тобто залишкову (випадкову) варіацію. Ця сума квадратів відхилень характеризує варіацію, зумовлену дією випадкових факторів, не врахованих у досліді.

Фактичне значення критерію F у цьому разі обчислюють як відношення дисперсії групових середніх (факторної) у варіантах досліду до залишкової дисперсії.

Оцінюючи відношення факторної дисперсії до залишкової, можна з певною ймовірністю зробити загальний висновок про вірогідність (відмову від гіпотези про випадкове варіювання) чи невірогідність (прийняття гіпотези) впливу факторної ознаки на результативну. Якщо дисперсійний аналіз підтверджує вірогідність розбіжностей між середніми в цілому у досліді, то цей загальний висновок можна доповнити оцінкою істотності різниці між груповими середніми. Таке попарне порівняння середніх дає змогу визначити, які варіанти

суттєво відрізняються від контролю і, отже, їх можна рекомендувати для впровадження у виробництво.

Для того, щоб визначити істотність різниці між будь-якими парами середніх, обчислюють середню помилку вибіркових середніх на основі залишкової дисперсії.

При однаковій чисельності груп середню помилку різниці двох середніх визначають за формулою

$$m_{1-2} = S_{зал}\sqrt{\frac{2}{n}},$$

де $S_{зал}$ — середнє квадратичне відхилення, зумовлене залишковою варіацією.

При різній кількості спостережень (неоднаковій чисельності груп) середню помилку різниці середніх обчислюють за формулою середньої гармонічної простої

$$m_{1-2} = S_{зал}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = S_{зал}\sqrt{\frac{2}{\overline{n}}},$$

де $\overline{n}$ — середня чисельність груп.

У нашому прикладі середня помилка різниці середніх між будь-якою парою вибірок дорівнюватиме

$$m_{1-2} = S_{зал}\sqrt{\frac{2}{n}} = 13{,}08 \cdot 0{,}707 = 9{,}25.$$

Виходячи з рівня ймовірності, прийнятого при дисперсійному аналізі (0,95), за таблицями t-критерію Стьюдента (дод. 2) визначають критичне значення t. При цьому беруть кількість ступенів свободи залишкової дисперсії.

У нашому прикладі помилка досліду ($S^2_{зал}$) має 15 ступенів свободи. За таблицею Стьюдента 15 ступеням свободи та ймовірності 0,95 відповідає t = 2,1315. Отже гранична помилка різниці середніх становитиме

$$\varepsilon_p = tm_{1-2} = 2{,}1315 \cdot 9{,}25 = 19{,}7ц\ .$$

Графічну помилку різниці середніх у статистиці називають **найменшою істотною різницею (НІР)**. Вона показує величину можливих випадкових коливань при певному рівні ймовірності. Якщо різниця врожайності порівнюваних варіантів перевищує розмір граничної помилки, то її можна вважати істотною. У табл. 46 наведено істотність розбіжностей між окремими варіантами і контролем.

Таблиця 46 *Істотність різниці між окремими варіантами і контролем*

Варіант досліду	Середня врожайність, ц/га	Різниця в урожайності відносно контролю, ц	Висновок про істотність різниці
Контроль	361	—	—
20	386	25	істотна
40	396	35	“
60	417	56	“
80	414	53	“
100	416	55	“

Отже, всі різниці середніх істотні, оскільки перевищують граничну помилку різниці середніх, тобто перевищують найменшу

істотну різницю. Відмінності в урожайності пояснюються дією досліджуваного фактора (добрив), а не випадковими коливаннями. Так само можна порівнювати і окремі варіанти досліду.

Спеціальні дослідження показують, що при трьох і більше варіантах у досліді оцінка відмінностей між середніми при використанні критерію t-Стьюдента дає більшу кількість істотних різниць. У зв’язку із цим пропонується інший спосіб попарного порівняння середніх. Цей метод грунтується на порівнянні попарних різниць з величиною $D = Qm$, де m — середня помилка середніх $\left(m = \sqrt{\frac{S^2_{зал}}{n}}\right)$, а Q — критерій Тьюкі (дод. 3). Величина Q залежить від кількості ступенів свободи залишкової дисперсії та кількості варіантів у досліді. Різниця між парами середніх, які перевищують величину D, є істотними при відповідному рівні ймовірності.

5.3 Дисперсійний аналіз при групуванні даних за двома ознаками

У статистичній практиці часто доводиться мати справу з багатофакторними дослідами, в яких одночасно вивчається вплив на результативну ознаку двох і більше факторів. У цьому разі виникає потреба в оцінці вірогідності впливу не тільки кожного фактора окремо, а й результату їх взаємодії. Цей результат визначають як різницю між ефектом впливу двох групувальних ознак і сумою ефектів кожної факторної ознаки, взятої окремо.

Схема аналізу при групуванні даних за двома ознаками така сама, як і при групуванні даних за однією ознакою. Проте розкладання варіації результативної ознаки на складові частини при двофакторних дослідах дещо складніше.

Припустімо, що в досліді вивчається одночасний вплив двох факторів (*А* і *Б*) на результативну ознаку. У незалежних вибірках, тобто в таких, які сформовані у випадковому порядку, джерелами варіювання результативної ознаки будуть перший (*А*) і другий (*Б*) фактори, їх можлива взаємодія та випадкове варіювання

$$W_{заг} = W_A + W_Б + W_{АБ} + W_{зал.}$$

Проте у польових дослідах із систематичним розміщенням варіантів принцип випадкового відбору спостережень у вибірку часто порушується. Причиною цього є неоднорідність окремих частин дослідної ділянки (різні рельєф, механічний склад грунту, родючість тощо). При розміщенні варіантів у певному порядку у всіх повторностях деякі варіанти можуть потрапити на більш родючі ділянки і їх урожайність буде завищеною.

Щоб уникнути цього, дослідну ділянку спочатку розподіляють на окремі частини (блоки), які мають деяку відмінність (у рельєфі, грунті). Після цього кожний варіант досліду у випадковому порядку розподіляють по всіх блоках, завдяки чому досягається рівноцінне

представлення ділянок за градаціями досліду. Отже, спостереження за одним варіантом тісно пов'язані зі спостереженнями за іншим варіантом, розміщеним в одному блоці. Цей спосіб побудови досліду, коли спочатку ділянку поділяють на блоки з відносно вирівняними умовами, а досліджувані варіанти розміщують усередині блоку у випадковому порядку, називають **способом рендомізованих блоків**.

Якщо групи та підгрупи сформовані як залежні вибірки, то джерелами варіації будуть перший (*А*) і другий (*Б*) фактори, їх можлива взаємодія (*АБ*), ділянки та залишкове варіювання

$$W_{заг.} = W_A + W_Б + W_{АБ} + W_{діл.} + W_{зал.}$$

Використання дисперсійного аналізу при двофакторному досліді розглянемо на такому прикладі (табл. 47).

Таблиця 47 *Урожайність озимої пшениці, ц/га*

Попередник	Мінеральні добрива	Урожайність на ділянках				$\sum x_j$	$\bar{x}_j$
		1	2	3	4		
Чорний пар	$N_{60}K_{60}$ (фон)	51,4	47,2	49,8	47,0	195,4	48,8
	фон + P_{60}	56,8	53,7	52,9	49,2	212,6	53,2
	фон + P_{90}	62,8	61,2	60,7	59,2	243,9	61,0
Горох	$N_{60}K_{60}$ (фон)	49,4	46,6	45,6	43,8	185,4	46,3
	фон + P_{60}	52,4	50,8	49,4	48,0	200,6	50,1
	фон + P_{90}	56,6	57,1	54,2	53,9	221,8	55,4

Продовження таблиці 47

Попередник	Мінеральні добрива	Урожайність на ділянках				$\sum x_j$	$\bar{x}_j$
		1	2	3	4		
	$\sum x_i$	329,4	316,6	312,6	301,1	1259,7	52,5
	$\bar{x}_i$	54,9	52,8	52,1	50,2	52,5	—

У нашому прикладі дослід закладений способом рендомізованих блоків. Загальну суму квадратів відхилень урожайності озимої пшениці від середньої в досліді в цілому розкладаємо на п'ять частин відповідно до джерел варіації

$$W_{заг} = W_{поп} + W_{доб} + W_{вз} + W_{діл} + W_{зал.}$$

Отже, варіація врожайності озимої пшениці в умовах досліду визначається попередниками, фосфорними добривами, взаємодією факторів, ділянками та випадковими причинами.

Суми квадратів відхилень при дисперсійному аналізі даних двофакторного досліду обчислюють так: спочатку із загальної варіації виділяють варіацію, що залежить від двох факторів, а також варіацію, пов'язану з ділянками, і залишкову варіацію:

$$W_{заг} = W_{поп+доб} + W_{діл} + W_{зал.}$$

Потім суму квадратів відхилень, пов'язану з впливом факторів, поділяють на варіацію, що залежить від попередників, добрив та взаємодії факторів:

$$W_{поп+доб} = W_{поп} + W_{доб} + W_{вз.}$$

У нашому прикладі загальна кількість спостережень $N = 24$, кількість груп по попередниках $K = 2$, кількість підгруп за дозами фосфорних добрив $l = 3$ і ділянках $n = 4$ ($N = Kln = 2 \cdot 3 \cdot 4 = 24$).

Щоб спростити розрахунки, зменшимо вихідні дані на постійну величину $a = 50$. Результати цих перетворень наведено в табл. 48.

Для перевірки правильності розрахунків у табл. 48 до суми відхилень від постійної величини ($\sum y_{ij} = 59,7$) додаємо добуток кількості ділянок ($N = 24$) і постійної величини ($a = 50$):

$$\sum x_{ij} = \sum y_{ij} + aN = 59,7 + 50 \cdot 24 = 1259,7.$$

Таблиця 48 *Відхилення від умовного нуля ($y = x - a$, $a = 50$)*

Попередник	Мінеральні добрива	Урожайність на ділянках				$\sum y_j$
		1	2	3	4	
Чорний пар	$N_{60}K_{60}$ (фон)	1,4	–2,8	–0,2	–3,0	–4,6
	фон + P_{60}	6,8	3,7	2,9	–0,8	12,6
	фон + P_{90}	12,8	11,2	10,7	9,2	43,9
Горох	$N_{60}K_{60}$ (фон)	–0,6	–3,4	–4,4	–6,2	–14,6
	фон + P_{60}	2,4	0,8	0,6	–2,0	0,6
	фон + P_{90}	6,6	7,1	7,2	3,9	21,8
	$\sum y_i$	29,4	16,6	12,6	1,1	59,7

Щоб визначити суми квадратів відхилень середніх у групах і підгрупах від загальної середньої, а також середніх на ділянках від загальної середньої і залишкової суми квадратів відхилень, піднесемо до квадрата всі індивідуальні відхилення від умовного нуля, їх суми та загальний підсумок (табл. 49).

Таблиця 49 *Визначення сум квадратів відхилень*

Попередник	Мінеральні добрива	Урожайність на ділянках				Сума квадратів $\sum y_j^2$	Квадрат суми $(\sum y_j)^2$
		1	2	3	4		
Чорний пар	$N_{60}K_{60}$ (фон)	1,96	7,84	0,04	9,00	18,84	21,16
	фон + P_{60}	46,24	13,69	8,41	0,64	68,98	158,76
	фон + P_{90}	163,84	125,44	114,49	84,64	488,41	1927,21
Горох	$N_{60}K_{60}$ (фон)	0,36	11,56	19,36	38,44	69,72	213,16
	фон + P_{60}	5,76	0,64	0,36	4,00	10,76	0,36
	фон + P_{90}	43,56	50,41	17,64	15,21	126,82	475,24
Сума квадратів	$\sum y_i^2$	261,72	209,58	160,30	151,93	783,53	2795,89
Квадрат сум	$(\sum y_i)^2$	864,36	275,56	158,76	1,21	1299,89	3564,09

Загальна сума квадратів відхилень становитиме

$$W_{заг} = \sum y_{ij}^2 - \frac{(\sum y_{ij})^2}{N} = 783{,}53 - \frac{3564{,}09}{24} = 635{,}03.$$

Сума квадратів відхилень, зумовлена впливом попередників і фосфорних добрив, дорівнюватиме

$$W_{поп+доп} = \frac{\sum(\sum y_j)^2}{n} - \frac{(\Delta y_{ij})^2}{N} = \frac{2795{,}89}{4} - \frac{3564{,}09}{34} = 550{,}47.$$

Слід урахувати, що при визначенні $W_{поп+доб}$ суму $\sum(\sum y_j)^2$ ділять на кількість ділянок, оскільки $\sum y_j$ є підсумком чотирьох ділянок.

Сума квадратів відхилень, зумовлена ділянками, дорівнюватиме

$$W_{діл} = \frac{\sum(\sum y_i)^2}{Kl} = \frac{(\sum y_{ij})^2}{N} = \frac{1299{,}89}{6} - \frac{3564{,}09}{24} = 68{,}15.$$

При визначенні $W_{діл}$ суму $\sum(\sum y_j)^2$ ділять на кількість варіантів досліду, оскільки сума $\sum y_i$ є підсумком шести варіантів.

Залишкову суму квадратів визначають як різницю між загальною сумою квадратів і сумами квадратів відхилень, зумовлених впливом обох факторів і ділянок:

$$W_{зал} = W_{заг} - W_{поп+доб} - W_{діл} = 635{,}03 - 550{,}47 - 68{,}15 = 16{,}41.$$

Для кожної з обчислених сум відхилень визначаємо кількість ступенів свободи варіації:

$$v_{заг} = N - 1 = 24 - 1 = 23;$$

$$v_{поп+доб} = Kl - 1 = 6 - 1 = 5;$$

$$v_{діл} = n - 1 = 4 - 1 = 3;$$

$$v_{зал} = (N - 1) - (Kl - 1) - (n - 1) = 23 - 5 - 3 = 15.$$

Далі вичленимо із суми квадратів відхилень, яка визначає коливання врожайності за рахунок обох факторів та взаємодії, складові її частини

$$W_{поп+доб} = W_{поп} + W_{доб} + W_{вз}.$$

Щоб їх розрахувати, за даними останньої графи табл. 49 складемо нову таблицю відхилень, у якій виділимо попередники та дози фосфорних добрив, причому відмінності в ділянках поєднаємо (табл. 50). Позначимо підсумкові графи та рядки таблиці для першого фактора (попередники) — через *А*, для другого (дози добрив) — через *Б*.

Таблиця 50 *Суми відхилень для визначення ефектів А, Б і взаємодії АБ*

Попередник *А*	Мінеральні добрива *Б*			Сума *А*
	фон	фон + P_{60}	фон + P_{90}	
Чорний пар	–4,6	12,6	43,9	51,9
Горох	–14,6	0,6	21,8	7,8
Сума *Б*	–19,2	13,2	65,7	59,7

Піднесемо до квадрата дані табл. 50 (табл. 51).

Таблиця 51 *Квадрати сум відхилень для визначення ефектів А, Б і взаємодії АБ*

Попередник *А*	Мінеральні добрива *Б*			Сума квадратів	Квадрат сум *А*
	фон	фон + P_{60}	фон + P_{90}		
Чорний пар	21,16	158,76	1927,21	2107,13	2693,61
Горох	213,16	0,36	475,24	688,76	60,84
Сума квадратів	234,32	159,12	2402,45	2795,89	2754,45
Квадрат сум *Б*	368,64	174,24	4316,49	4859,37	3564,09

Сума квадратів табл. 51 повторює квадрат суми табл. 50, на основі якої була обчислена варіація врожайності, зумовлена дією обох факторів, $W_{поп+доб}$ = 550,47 з 5 ступенями свободи варіації. Ця загальна сума квадратів відхилень складається із сум варіації, зумовленої попередниками, добривами та їх взаємодією:

$$W_{поп+доб} = W_{поп} + W_{доб} + W_{вз}.$$

Сума квадратів відхилень, зумовлена попередниками

$$W_{поп} = \frac{\sum A^2}{nl} - \frac{\left(\sum y_{ij}\right)^2}{N} = \frac{2754,45}{12} - \frac{3564,09}{24} = 81,04.$$

Сума квадратів відхилень, зумовлена добривами

$$W_{доб} = \frac{\sum Б^2}{nK} - \frac{\left(\sum y_{ij}\right)^2}{8} = \frac{4859,37}{8} - \frac{3564,09}{24} = 458,92.$$

Сума квадратів відхилень, зумовлена взаємодією попередників і добрив

$$W_{вз.} = W_{поп+доб} - W_{поп} - W_{доб} = 550{,}47 - 81{,}04 - 458{,}92 = 10{,}51.$$

Отже, перший етап розподілу загальної суми квадратів відхилень дав такі результати:

$$W_{заг} = W_{поп+доб} + W_{діл} + W_{зал};$$
$$635{,}03 = 550{,}47 + 68{,}15 + 16{,}41;$$

Таблиця 52 *Аналіз дисперсій*

Джерела варіації	Сума квад-ратів відхилень W	Ступені свободи варіації v	Дисперсія S^2	Відношення дисперсій	
				$F_ф$	F_T
Попередники	81,04	1	81,04	74,35	4,54
Добрива	458,92	2	229,46	210,51	3,68
Взаємодія факторів	10,51	2	5,25	4,82	3,68
Ділянки	68,15	3	22,72	20,84	3,29
Залишкова варіація	16,41	15	1,09	1	—
Загальна варіація	635,03	23	—	—	—

другий етап розподілу:

$$W_{поп+доб} = W_{поп} + W_{доб} + W_{вз};$$
$$550{,}47 = 81{,}04 + 458{,}92 + 10{,}51;$$

у цілому:

$$W_{заг} = W_{поп} + W_{доб} + W_{вз.} + W_{діл} + W_{зал};$$
$$635{,}03 = 81{,}04 + 458{,}92 + 10{,}51 + 68{,}15 + 16{,}41.$$

Для того щоб обчислити дисперсію, слід визначити ступені свободи варіації для кожної суми квадратів відхилень. Відомо, що ступені свободи варіації для загальної суми квадратів становлять $V = 23$, для суми квадратів відхилень, зумовленої впливом обох факторів, $V = 5$, для суми квадратів відхилень, пов'язаної з ділянками, $V = 3$ і для залишкової — $V = 15$.

Розподілимо ступені свободи варіації обох факторів ($V_{поп+доб} = 5$) на три компоненти, які становлять цю суму квадратів відхилень.

Для $W_{поп}$ з двома середніми кількість ступенів свободи становитиме: $V_{поп} = K - 1 = 2 - 1 = 1$; для $W_{доб}$ з трьома середніми $V_{доб} = 1 - 1 = 3 - 1 = 2$; для суми квадратів відхилень, зумовленої взаємодією факторів — $V_{вз} = V_{поп + доб} - V_{поп} - V_{доб} = 5 - 1 - 2 = 2$.

Обчислення дисперсій та їх оцінку при рівні ймовірності $P = 0{,}95$ наведено в табл. 52.

Перевіримо гіпотезу про випадковий характер розбіжностей між середньою врожайністю озимої пшениці за досліджуваними факторами та висновок по цій гіпотезі зробимо з ймовірністю помилитися лише в 5 випадках із 100 ($P = 0{,}95$).

Порівняння фактичних відношень порівнюваних дисперсій з їх максимальною величиною при надійній ймовірності показує, що у всіх випадках фактичне значення критерію F перевищує його табличне значення ($F_ф > F_T$). Отже, відмінності в урожайності озимої пшениці під впливом попередників і доз фосфорних добрив вірогідні. Імовірний також і ефект взаємодії обох факторів. Звідси висунута гіпотеза не узгоджується з фактичними даними, розбіжності в середній врожайності озимої пшениці під впливом досліджуваних факторів не пов'язані з випадковим варіюванням.

Висновок про вірогідність відмінностей, зумовлених дією окремих факторів і їх взаємодією, дає змогу попарно порівняти середні по кожному фактору і середні, які відображують загальну взаємодію факторів.

Середня помилка різниці середніх дорівнюватиме

$$m_{1-2} = S_{зал}\sqrt{\frac{2}{n}} = 1{,}05 \cdot 0{,}707 = 0{,}7 \ ц/га.$$

При ймовірності $P = 0{,}95$ і 15 ступенях свободи варіації залишкової дисперсії за таблицею Стьюдента (додаток 2) нормоване відхилення $t = 2{,}1315$. Гранична помилка різниці середніх становитиме

$$\varepsilon_p = tm_{1-2} = 2{,}1315 \cdot 0{,}7 = 1{,}5 ц/га.$$

Це означає, що всі порівнювані середні, різниця між якими більша ніж 1,5 ц, слід вважати вірогідними.

5.4 Дисперсійний аналіз альтернативних ознак

Зазначені вище принципи дисперсійного аналізу кількісних ознак можна застосувати також при порівнянні кількох вибірок якісних (альтернативних) ознак. **Альтернативними** називають такі ознаки, які мають одні одиниці досліджуваної сукупності і не мають інших одиниць. Кількісно варіацію альтернативної ознаки виражають двома значеннями: наявність ознаки позначають через 1, а відсутність — через 0.

Суму квадратів загальної дисперсії альтернативної ознаки обчислюють за формулою

$$W_{заг} = \sum n - \frac{(\sum n)^2}{\sum m},$$

де n — кількість одиниць досліджуваного явища, які мають ознаку; m — загальна кількість одиниць в окремих вибірках.

Суму квадратів групової дисперсії (варіантів досліду) визначають за формулою

$$W_{гр} = \sum np - \frac{(\sum n)^2}{\sum m},$$

де p — частка ознаки $\left(p = \frac{n}{m}\right)$.

Суму квадратів залишкової дисперсії обчислюють як різницю між сумами квадратів загальної та групової дисперсій.

Кількість ступенів свободи загальної дисперсії визначають як $\sum m - 1$, групової — $K - 1$ (де K – кількість груп), залишкової — $(\sum m - 1) - (K - 1)$.

Дисперсійний аналіз альтернативної ознаки розглянемо на такому прикладі (табл. 53).

Таблиця 53 *Польова схожість* насіння *кукурудзи і розрахунки для обчислення сум квадратів*

Показник	Польова схожість насіння кукурудзи за зразками						Сума
	1	2	3	4	5	6	
Висіяно насіння	84	81	86	78	72	83	484
Зійшло насінин	56	65	73	43	67	42	346
$P = n:m$	0,67	0,80	0,85	0,55	0,93	0,5	–
N_p	37,52	52,00	62,05	23,65	62,31	21,0	258,53

На підставі даних табл. 53 обчислюємо суми квадратів

$$W_{заг} = \sum n - \frac{(\sum n)^2}{\sum m} = 346 - \frac{346^2}{484} = 98{,}65;$$

$$W_{гр} = \sum np - \frac{(\sum n)^2}{\sum m} = 258{,}53 - \frac{346^2}{484} = 11{,}18;$$

$$W_{зал} = W_{заг} - W_{гр} = 98{,}65 - 11{,}18 = 87{,}47.$$

Для кожної з обчислених сум квадратів визначаємо кількість ступенів свободи варіації:

$$V_{заг} = \sum m - 1 = 44 - 1 = 483;$$
$$V_{гр} = K - 1 = 6 - 1 = 5;$$
$$V_{зал} = (\sum m - 1) - (K - 1) = 483 - 5 = 478.$$

Розрахунок дисперсій та їх оцінку при рівні ймовірності $p = 0{,}95$ наведено в табл.54.

Таблиця 54 *Аналіз дисперсій*

Джерела варіації	Сума квадратів	Кількість ступенів свободи	Дисперсія	Відношення дисперсій	
				F_ϕ	F_T при ймовірності 0,95
Варіанти	11,18	5	2,24	12,44	2,21
Залишкова варіація	87,47	478	0,18	1	—
Загальна варіація	98,65	483	—	—	—

Як показують дані табл. 54, фактичне відношення між порівнюваними дисперсіями значно перевищує табличне. Це дає змогу зробити висновок про наявність вірогідних відмінностей між зразками кукурудзи.

Висновок про вірогідність відмінностей, зумовлених дією факторної ознаки, дає можливість попарно порівняти частки за окремими зразками. Середню помилку різниці часток визначають за формулою

$$m_{1-2} = S_{зал}\sqrt{\frac{2}{n}} = 0{,}42 \cdot 0{,}065 = 0{,}0273.$$

При ймовірності $P = 0{,}95$ і 478 ступенях свободи варіації залишкової дисперсії за таблицею Стьюдента (додаток 2) нормоване відхилення $t = 1{,}96$.

Гранична помилка різниці часток становитиме

$$\varepsilon_p = tm_{1-2} = 1{,}96 \cdot 0{,}0273 = 0{,}054.$$

Це означає, що при ймовірності $P = 0,95$ неістотно відрізняються за польовою схожістю насіння кукурудзи тільки пари зразків 2 – 3 і 4 – 6.

Контрольні питання

1. *Що таке дисперсійний аналіз?*
2. *З якою метою застосовують дисперсійний аналіз?*
3. *Що таке систематична та випадкова варіації?*
4. *Що таке кількість свободи варіації?*
5. *В чому суть критерію F-розділу?*
6. *Яка принципова схема дисперсійного аналізу?*
7. *Яка послідовність дисперсійного аналізу при групуванні даних за однією ознакою?*
8. *Яка послідовність дисперсійного аналізу при групуванні даних за двома ознаками?*
9. *Як здійснюється дисперсійний аналіз альтернативних ознак?*

Розділ 6
КОРЕЛЯЦІЙНИЙ АНАЛІЗ

6.1 Види взаємозв'язків і поняття про кореляційний аналіз

Усі явища, що існують у природі та суспільстві, перебувають у взаємозалежності і взаємозумовленості. Так, продуктивність праці залежить від розвитку і впровадження у виробництво досягнень науково-технічного прогресу, введення в дію основних фондів — від обсягу інвестицій в основний капітал, фондомісткість видобутку вугілля — від глибини розробки вугільних пластів, трудомісткість цукру — від потужності цукрових заводів, урожайність сільськогосподарських культур — від обсягу та інтенсивності агротехнічних заходів, продуктивність тварин — від рівня та якості годівлі, захворюваність населення — від екологічного стану навколишнього середовища.

За ступенем залежності одного явища від іншого розрізняють два види зв'язку: функціональний (повний) і стохастичний (неповний або статистичний).

Функціональним називають такий зв'язок, при якому кожному значенню факторної ознаки x (аргументу), що характеризує певне явище, в усіх випадках відповідає одне або кілька значень результативної ознаки y (функції). Прикладом такого зв'язку є залежність між радіусом і довжиною кола, стороною та площею квадрата. Функціональна залежність виявляється у кожному окремому випадку абсолютно точно і виражається за допомогою аналітичних формул. Знаючи x, можна в кожному окремому випадку точно визначити значення y. У соціально-економічних явищах до функціонального типу належать зв'язки між показниками — адитивні ($y = x_1 + x_2$) або мультиплікативні ($y = x_1 \cdot x_2$; $y = x_1 / x_2$), а також залежність середніх величин від структури сукупності.

На відміну від функціонального, стохастичний зв'язок неоднозначний. При ньому кожному значенню ознаки відповідає певна множина значень ознаки y, які утворюють так званий умовний розподіл. Якщо умовні розподіли замінюють одним параметром — середнім значенням $\overline{y}_i$ то такий зв'язок називають **кореляційним**.

Отже, кореляційний зв'язок є різновидом стохастичного і виявляється в зміні середніх умовних розподілів.

Кореляційний аналіз — це метод визначення та кількісної ознаки взаємозалежностей між статистичними ознаками, що характеризують окремі соціально-економічні явища та процеси.

При **кореляційному зв'язку** немає суворої відповідності між значеннями залежних ознак: кожному певному значенню аргументу (факторної ознаки) відповідає кілька різних значень функції (результативної ознаки). Такий, наприклад, зв'язок між потужністю цукрових заводів і трудомісткістю виробництва цукру. При однакових значеннях потужності заводів затрати праці на виробництво 1 тонни цукру будуть різні, тому що трудомісткість цукру залежить не тільки від потужності заводів, а й від багатьох інших факторів: технічного стану та рівня використання виробничого обладнання, якості і своєчасності переробки сировини, кваліфікації робочої сили тощо. Зв'язок між потужністю заводів і трудомісткістю цукру виявиться лише тоді, коли взяти достатню кількість спостережень і порівняти середні значення факторної та результативної ознак. Подібний зв'язок існує між дозами унесених добрив і урожайністю сільськогосподарських культур. При тих самих дозах добрив урожійність на різних ділянках буде неоднакова, тому що урожайність залежить не тільки від добрив, а й від багатьох інших факторів: сорту, своєчасності і якості виконання агротехнічних заходів, кількості онадів, температури тощо. Зв'язок між добривами і урожайністю виявиться лише тоді, коли узяти достатню кількість спостережень і порівняти середні значення факторної та результативної ознак.

На відміну від функціонального зв'язку кореляційний зв'язок виявляється не в кожному окремому випадку, а при великій кількості спостережень під час порівняння середніх значень взаємозалежних ознак. Він грунтується на законі великих чисел, який виявляється у масовому процесі як тенденція до зростання чи зменшення результативної ознаки залежно від відповідної зміни факторної ознаки. З математичного погляду кореляційна залежність —

це функціональне співвідношення тільки між середніми значеннями досліджуваних ознак.

За напрямом зв'язок між корелюючими величинами може бути прямим та оберненим. При **прямому зв'язку** зміна факторної ознаки зумовлює зміну результативної ознаки в тому самому напрямі (наприклад, зв'язок між унесенням добрив та урожайністю сільськогосподарських культур, рівнем годівлі та продуктивністю тварин, рівнем механізації виробничих процесів та продуктивністю праці).

Якщо із збільшенням факторної ознаки результативна ознака зменшується чи, навпаки, із зменшенням факторної ознаки результативна зростає, то такий зв'язок називають **оберненим** (наприклад, зв'язок між урожайністю та собівартістю продукції, собівартістю продукції та рентабельністю виробництва, продуктивністю праці та собівартістю продукції).

За формою зв'язку розрізняють прямолінійні та криволінійні кореляційні залежності. **Прямолінійний кореляційний зв'язок** характеризується рівномірним зростанням або зменшенням результативної ознаки під впливом відповідної зміни факторної ознаки. Аналітично його визначають за рівнянням прямої лінії.

При **криволіиійному кореляційному зв'язку** однаковим змінам середніх значень факторної ознаки відповідають різні зміни середніх значень результативної ознаки. Аналітично криволінійний зв'язок визначають за рівнянням кривої лінії.

Залежно від кількості досліджуваних ознак розрізняють парну (просту) та множинну кореляцію. При **парній кореляції** аналізують зв'язок між факторною та результативною ознаками; при **множинній кореляції** — залежність результативної ознаки від двох та більше факторних ознак.

За допомогою кореляційного аналізу розв'язують такі завдання: виявляють навявність та вибір форми зв'язку результативної ознаки з одним або комплексом факторів; кількісно оцінюють зміни залежної величини від факторів, що впливають на неї; встановлюють тісноту зв'язку результативного показника з одним фактором чи їх

комплексом; аналізують загальний обсяг варіації залежної величини та визначають вплив окремих факторів у цьому варіюванні; статистично оцінюють вибіркові показники кореляційного зв'язку.

Схематично кореляційний аналіз складається з таких послідовних стадій:

1) встановлення та відбору найбільш істотних ознак для аналізу;
2) визначення напряму та форми зв'язку результативного та факторних показників та вибору типу математичного рівняння для аналізу існуючих зв'язків;
3) розрахунку характеристик кореляційної залежності;
4) статистичної оцінки вибіркових показників зв'язку.

Для того щоб правильно застосувати кореляційні методи, потрібно насамперед глибоко вивчити суть взаємозв'язків соціально-економічних явищ. Ці методи не виявляють причин виникнення зв'язків між окремими явищами та характеру їх взаємодії. Характер взаємозв'язків та закономірностей розвитку економічних процесів встановлюють за допомогою теоретичного аналізу. Кореляційний метод включає кількісну оцінку взаємозалежностей між статистичними ознаками, що характеризують досліджувані явища.

Нійбільш істотні ознаки для аналізу відбирають логіко-теоретичним шляхом залежно від змісту співвідношення результативної та факторної ознак. При цьому важливу роль відіграє попередній аналіз досліджуваного явища, який є основою для визначення завдання кількісного вивчення зв'язку. Попередній аналіз передбачає порівняння по можливості взаємозалежних статистичних рядів, побудову таблиць розподілу чисельностей за двома ознаками та їх графіків, застосування простих та комбінованих групувань за факторними або результативними ознаками.

Напрям та форма зв'язку та вибір типу математичного рівняння найбільш чітко визначають взаємозалежність факторної та результативної ознак. Напрям та форму зв'язку встановлюють за допомогою статистичних групувань, а також графіків, побудованих у системі прямокутних координат на основі емпіричних даних.

Графічне зображення статистичних показників дає наочне уявлення про наявність зв'язку між досліджуваними ознаками. При побудові графіка на горизонтальній осі відкладають значення факторної ознаки (x), а на вертикальній — значення результативної ознаки (y). Відклавши на перетині відповідних значень x та y точки, дістають кореляційне поле.

За характером розміщення точок на кореляційному полі роблять висновок про напрям та форму зв'язку. Якщо точки безладно розміщені по всьому полю, то це свідчить про те, що залежності між досліджуваними ознаками немає. Якщо точки концентруються навколо уявної осі, напрям якої від нижнього лівого кута до верхнього правого, то зв'язок між результативною та факторною ознакою прямий. Коли ж точки концентруються навколо уявної осі, напрям якої з верхнього лівого кута до нижнього правого, то існує обернений зв'язок між ознаками. Характер розподілу точок по кореляційному полю вказує і на наявність прямолінійної чи криволінійної залежності між факторною та результативною ознаками.

За допомогою графіка співвідношення досліджуваних ознак роблять висновок про вибір типу математичного рівняння для кількісної оцінки зв'язку. Рівняння, за допомогою яких визначають статистичний зв'язок між корелюючими величинами, називають **рівняннями регресії (кореляційними рівняннями)**, а лінії, побудовані на їх основі, — **лініями регресії**.

6.2 Парна (проста) прямолінійна кореляція

Залежно від форми зв'язку між факторною та результативною ознаками вибирають тип математичного рівняння. Прямолінійну форму зв'язку визначають за рівнянням прямої лінії

$$y_x = a_0 + a_1 x,$$

де y_x — теоретичні (обчислені за рівнянням регресії) значення результативної ознаки; a_0 — початок відліку, або значення y_x при умові, що $x = 0$; a_1 — коефіцієнт регресії (коефіцієнт пропорційності), який показує, як змінюється y_x при кожній зміні x на одиницю; x — значення факторної ознаки.

При прямому зв'язку між корелюючими ознаками коефіцієнт регресії a_1 матиме додатне значення, при зворотному — від'ємне.

Параметри a_0 і a_1 рівняння регресії обчислюють способом найменших квадратів. Суть цього способу в знаходженні таких параметрів рівняння зв'язку, при яких залишкова сума квадратів відхилень фактичних значень результативної ознаки (y) від її теоретичних (обчислених за рівнянням зв'язку) значень (y_x) буде мінімальною:

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

Спосіб найменших квадратів зводиться до складання та розв'язання системи двох рівнянь з двома невідомими:

$$\sum y = na_0 + a_1 \sum x;$$
$$\sum xy = a_0 \sum x + a_1 \sum x^2,$$

де n — кількість спостережень.

Розв'язавши цю систему рівнянь у загальному вигляді, матимемо формули для визначення параметрів a_0 і a_1:

$$a_0 = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \sum x \sum x}; \quad a_1 = \frac{n\sum xy - \sum x \sum xy}{n\sum x^2 - \sum x \sum x}.$$

Обчислення показників зв'язку розглянемо на прикладі (табл. 55).

Таблиця 55 *Вихідні та розрахункові дані для обчислення параметрів рівняння зв'язку між унесенням добрив та урожайністю зернових культур*

Номер господарства	Унесено органічних і мінеральних добрив на 1 га посівів, ц.д.р., x.	Урожайність зернових культур, ц/га	Розрахункові величини			Теоретичні значення урожайності, ц/га
		y	xy	x^2	y^2	$y_x = 30,27 + 4,56x$
1	1,1	35	38,5	1,21	1225	35,3
2	1,9	38	72,2	3,61	1444	38,9
3	2,1	40	84,0	4,41	1600	39,8
4	1,4	36	50,4	1,96	1296	36,7
5	1,6	37	59,2	2,66	1369	37,6
6	1,6	39	62,4	2,56	1521	37,6
7	1,2	36	43,2	1,44	1296	35,7
8	1,9	39	74,1	3,61	1521	38,9
9	1,4	37	51,8	1,96	1369	36,6
10	2,1	40	84,0	4,41	1600	39,9
Разом	16,3	377	619,8	27,73	14241	377
У середньому	1,63	37,7	61,98	2,773	1424,1	37,7

Підставивши нідсумкові дані табл. 55 у наведені вище формули, дістанемо

$$a_0 = \frac{27{,}73 \cdot 377 - 16{,}3 \cdot 619{,}8}{10 \cdot 7{,}73 - 16{,}3 \cdot 16{,}3} = 30{,}27 \ ц/га;$$

$$a_1 = \frac{10 \cdot 619{,}8 - 16{,}3 \cdot 377}{10 \cdot 7{,}73 - 16{,}3 \cdot 16{,}3} = 4{,}56 \ ц/га.$$

Отже, рівняння кореляційного зв'язку між урожайністю зернових культур та унесенням органічних та мінеральних добрив матиме вигляд

$$y_x = 30{,}27 + 4{,}56x.$$

Економічний зміст цього рівняння такий: коефіцієнт регресії показує, що в досліджуваній сукупності господарств із збільшенням дози унесення добрив на 1 ц діючої речовини врожайність зернових культур зростає в середньому на 4,56 ц/га. Параметр a_0 (у нашому прикладі 30,27) як вільний член рівняння має тільки розрахункове значення та не інтерпретується.

Підставивши y рівняння регресії значення x, дістанемо теоретичні рівні врожайності зернових культур у кожному господарстві. Якщо сума теоретичних значень дорівнює сумі емпіричних значень $(\sum y_x = \sum y)$, то параметри рівняння визначені правильно.

Якщо параметри рівняння регресії обчислюють для згрупованих даних, то формули для їх визначення з урахуванням частот мають такий вигляд:

$$a_0 = \frac{\sum yf \sum x^2 f - \sum xyf \sum xf}{\sum f \sum x^2 f - \sum xf \sum xf};$$

$$a_1 = \frac{\sum f \sum xyf - \sum yf \sum xf}{\sum f \sum x^2 f - \sum xf \sum xf}.$$

Порядок обчислення параметрів a_0 і a_1 розглянемо на такому прикладі (табл. 56).

Таблиця 56 *Вихідні та розрахункові дані для обчислення параметрів рівняння зв'язку між змінним виробітком бурякозбиральних комбайнів та стажем роботи трактористів-машиністів*

Стаж роботи	Середина інтервалу	Кількість трактористів-машиністів	Середній змінний виробіток, тис.грн.	Розрахункові величини				Змінний виробіток, обчислений за рівнянням регресії
	x	f	y	xf	Yf	x^2f	xyf	$y_x = 1{,}65 + 0{,}09x$
До 3	1	5	1,52	5	7,6	5	7,6	1,74
3 – 7	5	6	2,37	30	14,2	150	71,0	2,10
7 – 11	9	8	2,41	72	19,3	648	173,7	2,46
11 – 15	13	7	2,67	91	18,7	1183	243,1	2,82
15 і більше	17	4	3,15	68	12,6	1146	214,2	3,18
Разом	—	30	—	266	72,4	3142	709,6	—

Підставивши дані табл. 56 у наведені вище формули, матимемо

$$a_0 = \frac{72{,}4 \cdot 3142 - 709{,}6 \cdot 266}{30 \cdot 3142 - 266 \cdot 266} = 1{,}65;$$

$$a_1 = \frac{30 \cdot 709{,}6 - 72{,}4 \cdot 266}{30 \cdot 3142 - 266 \cdot 266} = 0{,}09.$$

Рівняння регресії, яке характеризує залежність змінного виробітку бурякозбиральних комбайнів від стажу роботи трактористів-машиністів, має такий вигляд:

$$y_x = 1{,}65 + 0{,}09x.$$

Коефіцієнт регресії (0,09) показує, що при збільшенні стажу роботи трактористів-машиністів на один рік змінний виробіток на комбайн підвищується в середньому на 0,09 тис.грн.

Рівняння регресії використовують для прогнозування очікуваних рівнів результативних ознак при встановлених значеннях факторних ознак. Порівнюючи очікувані рівні з фактично досягнутими, визначають резерви виробництва.

Важливим завданням кореляційного аналізу є визначення тісноти зв'язку між корелюючими величинами. Кількісним показником тісноти прямолінійного зв'язку результату з одним фактором є коефіцієнт парної кореляції, який обчислюють за формулою

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y},$$

де r — лінійний коефіцієнт кореляції; σ_x — середнє квадратичне відхилення факторної ознаки; σ_y — середнє квадратичне відхилення результативної ознаки.

Якщо врахувати, що $\sigma_x = \sqrt{\overline{x^2} - \bar{x}^2}$, а $\sigma_y = \sqrt{\overline{y^2} - \bar{y}^2}$, то найбільш зручною формулою для визначення лінійного коефіцієнта кореляції є

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}.$$

При парній залежності коефіцієнт кореляції коливається від 0 до +1 при прямому зв'язку та від 0 до –1 — при зворотному зв'язку. Чим ближче коефіцієнт кореляції до ±1, тим тісніший зв'язок між у та х і, навпаки, чим ближче коефіцієнт кореляції до 0, тим слабший зв'язок між результативною та факторною ознаками. Якщо $r < 0{,}3$, зв'язку немає, якщо $r = 0{,}3 - 0{,}5$ – зв'язок слабкий, якщо $r = 0{,}5 - 0{,}7$ — зв'язок середній і якщо $r > 0{,}7$ – зв'язок тісний. Коефіцієнт кореляції має такий самий знак, як і коефіцієнт регресії у рівнянні зв'язку.

На підставі даних табл. 56 обчислимо коефіцієнт кореляційної залежності урожайності зернових культур від унесених добрив:

$$r = \frac{61{,}98 - 1{,}63 \cdot 37{,}7}{\sqrt{(2{,}773 - 1{,}63^2)(14{,}241 - 37{,}7^2)}} = 0{,}926.$$

Коефіцієнт кореляції показує, що між дозами унесених добрив та урожайністю зернових культур у досліджуваних господарствах існує тісний зв'язок.

Складніше визначити коефіцієнт кореляції на основі згрупованих даних за двома ознаками. Його обчислення розглянемо на такому прикладі (табл. 57).

У табл. 57 прийнято такі умовні позначення: $d_x = x - \bar{x}$ — відхилення стажу роботи окремих груп трактористів-машиністів від середнього стажу у всій сукупності — відхилення варіантів змінного виробітку від середнього змінного виробітку ($\bar{y} = 2{,}4$).

Таблиця 57 *Вихідні та розрахункові дані для обчислення коефіцієнта кореляційного зв'язку між змінним виробітком бурякозбиральних комбайнів та стажем роботи трактористів-машиністів*

Стаж роботи, років (середина інтервалу) x	Змінний виробіток комбайнів, га, y							Сума ряда f_x	xfx	d_x	d_x^2	$d_x^2 fx$	$d_x d_y f_{xy}$
	1,2	2,0	2,4	2,5	2,7	2,8	3,5						
1	3	2						5	5	–8	64	320	35.2
5		1	3	2				6	30	–4	16	96	0.8
9		1	2	5				8	72	0	0	0	0
13				2	3	2		7	91	4	16	112	7,6
17						2	2	4	68	8	64	256	24,0
Сума, fy	3	4	5	9	3	4	2	30	266	—	—	784	67,6
									$\bar{x} = 9$				
yf_y	3,6	8,0	12,0	22,5	8,1	11,2	7,0	72,4	$\bar{y} = 2,4$				
d_y	–1,2	–0,4	0	0,1	0,3	0,4	1,1	—					
d_y^2	1,44	0,16	0	0,01	0,09	0,16	1,21	—					
$d_y^2 f_y$	4,32	0,64	0	0,09	0,27	0,64	2,42	8,38					

Особливу увагу потрібно звернути на обчислення величини $d_x d_y f_{xy}$. Наведемо приклади їх розрахунку:

1-й рядок (–8)·(–1,2)·3=28,8
(–8)·(–0,4)·2=6,4

Разом = 35,2

2-й рядок (–4)·(–0,4)·1=1,6
(–4)·0·3=0
(–4)·0,1·2= –0,8

Разом = 0,8

Коефіцієнт кореляції визначають за формулою

$$r = \frac{d_x d_y f_{xy}}{\sqrt{\sum d_x^2 f_x \cdot \sum d_y^2 f_y}} = \frac{67,6}{\sqrt{784 \cdot 8,38}} = 0,834.$$

Обчислений коефіцієнт кореляції показує, що між змінним виробітком та стажем роботи трактористів-машиністів існує тісний зв'язок.

Якщо кількісні показники кореляційного аналізу визначено за вибірковими даними, то слід установити їх відповідність показникам зв'язку у генеральній сукупності. У зв'язку з цим статистично оцінюють коефіцієнти регресії та кореляції. При цьому обчислюють надійний інтервал випадкових коливань показників зв'язку в генеральній сукупності.

При кореляційному зв'язку середню помилку визначають на основі залишкової дисперсії. Квадрат середньої помилки вибіркового коефіцієнта регресії — це відношення скоригованої залишкової дисперсії до суми квадратів відхилень незалежної змінної:

$$m^2 = \frac{S^2}{\sum\left(x - \bar{x}\right)^2},$$

де m^2 — квадрат середньої помилки вибіркового коефіцієнта регресії; S^2 — скоригована залишкова дисперсія.

Скориговаиу залишкову дисперсію визначають за формулою

$$S^2 = \sigma^2 \frac{n}{n-K},$$

де σ^2 — залишкова дисперсія; n — чисельність вибірки; K — кількість параметрів рівняння регресії (при парній лінійній залежності $K = 2$).

Відповідно середня помилка коефіцієнта регресії

$$m = \sqrt{\frac{S^2}{\sum\left(x - \bar{x}\right)^2}}.$$

Граничну помилку вибіркового коефіцієнта регресії обчислюють за такою формулою:

$$\Delta = tm,$$

де t — нормоване відхилення, розмір якого встановлюють за спеціальними таблицями. При великих вибірках значення t знаходять за дод. 1, при малих вибірках — за дод.2.

Надійні межі коефіцієнта регресії у генеральній сукупності

$$A = a_1 \pm \Delta,$$

f фактичне значення нормованого відхилення t

$$t_\phi = \frac{a_1}{m}.$$

Якщо tф > tт при встановленому рівні значущості та відповідній кількості ступенів свободи варіації, то нульова гіпотеза відхиляється та робиться висновок про вірогідність вибіркового коефіцієнта регресії. Якщо ж tф < tт, то нульова гіпотеза приймається та робиться висновок про те, що значення вибіркового коефіцієнта регресії не є істотним, випадковим.

Розглянемо порядок визначення вірогідності вибіркового коефіцієнта регресії, що характеризує залежність урожайності зернових культур від доз унесених добрив: $y_x = 30{,}3 = 4{,}6x$. Використовуючи параметри цього рівняння і дані табл. 56 визначимо залишкову дисперсію:

$$\sigma^2_{зал.} = \frac{\sum y^2 - a_0 \sum y - a_1 \sum xy}{n} = \frac{1441 - 30{,}27 \cdot 377 - 4{,}56 \cdot 619{,}8}{10} = 0{,}29.$$

Обчислену залишкову дисперсію скоригуємо на втрату ступенів свободи варіації:

$$S^2 = \sigma^2_{зал.} \frac{n}{n-K} = 0{,}29 \cdot \frac{10}{10-2} = 0{,}28.$$

Визначимо суму квадратів відхилень незалежної змінної:

$$W = \sum\left(x - \bar{x}\right)^2 = \sum x^2 - n\bar{x}^2 = 27{,}73 - 10 \cdot 1{,}63^2 = 1{,}16.$$

Середня помилка коефіцієнта регресії

$$m = \sqrt{\frac{S^2}{\sum(x-\bar{x})^2}} = \sqrt{\frac{0{,}36}{1{,}16}} = 0{,}56.$$

Фактичне значення нормованого відхилення t

$$t_\phi = \frac{a_1}{m} = \frac{4{,}56}{0{,}56} = 8{,}143.$$

Табличне значення t (дод. 2) при рівні значущості 0,05 та 8 ступенях свободи варіації ($n - K = 10 - 2$) дорівнює 2,307.

Оскільки $t_\phi > t_{0,05}$ (8,143 > 2,307), то від нульової гіпотези про те, що немає зв'язку між урожайністю та унесенням добрив у генеральній сукупності ($a_i = 0$), слід відмовитися. Вибірковий коефіцієнт регресії $a_i = 4{,}56$ є істотним.

Визначимо граничну помилку вибіркового коефіцієнта регресії:

$$\Delta = tm = 2{,}307 \cdot 0{,}56 = 1{,}29 \text{ ц/га на 1 ц діючої речовини.}$$

Маємо такий інтервал, в якому із встановленим рівнем значущості знаходиться коефіцієнт регресії у генеральній сукупності,

$$A = a_1 \pm \Delta \; x \; 4{,}57 \pm 1{,}29, \text{ або } 3{,}27 \leq a_1 \leq 5{,}85.$$

Отже, з імовірністю помилитися лише у 5 випадках із 100 можна стверджувати, що коефіцієнт регресії, характеризуючи зв'язок урожайності зернових культур з унесенням добрив у генеральній сукупності, буде лежати в інтервалі від 3,27 до 5,86 ц/га з розрахунку на 1 ц діючої речовини органічних та мінеральних добрив.

Вибірковий коефіцієнт кореляції підлягає статистичній оцінці істотності лише тоді, коли вибіркова сукупність формується у випадковому порядку. Якщо ж принцип випадковості відбору одиниць

у вибіркову сукупність з будь-яких причин (наприклад, при типовому відборі) був порушений, то вибірковий коефіцієнт кореляції не можна використати для характеристики цієї ж величини у генеральній сукупності.

Середню помилку вибіркового коефіцієнта парної кореляції обчислюють за такою формулою:

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

де m_r — середня помилка коефіцієнта кореляції; r — коефіцієнт кореляції; n — вибіркова сукупність.

Цю формулу використовують тоді, коли чисельність вибіркової сукупності досить велика, а значення вибіркового коефіцієнта кореляції не наближається до одиниці.

Вірогідність коефіцієнта кореляції визначають за t-критерієм, який обчислюють як відношення коефіцієнта кореляції до його середньої помилки:

$$t_\phi = \frac{r}{m_r}.$$

Якщо t_ϕ перевищує табличне значення t при встановленому рівні значущості, то роблять висновок, що в генеральній сукупності зв'язок між досліджуваними ознаками вірогідності. Коли ж $t_\phi < t_T$, то коефіцієнт кореляції невірогідний.

Якщо чисельність вибірки невелика, а вибірковий коефіцієнт кореляції наближається до одиниці, оцінку його істотності здійснюють за допомогою **методу перетвореної кореляції**. Р.Фішер довів, що розподіл логарифмічної функції вибіркового лінійного коефіцієнта кореляції (z) наближається до кривої нормального розподілу навіть при невеликому обсязі вибірки та значенні r, близькому до одиниці.

Вибірковий коефіцієнт кореляції визначають за формулою

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

Перехід від r до z здійснюють за допомогою спеціальних таблиць (дод. 9).

Середня помилка z-розподілу залежить від обсягу вибірки. Її обчислюють за такою формулою:

$$m = \frac{1}{\sqrt{n-3}}.$$

У нашому прикладі (див. табл. 56) середня помилка z-розподілу

$$m = \frac{1}{\sqrt{10-3}} = \frac{1}{\sqrt{7}} = 0{,}378.$$

За дод. 9 знаходимо, що коефіцієнту кореляції 0,926 відповідає $z = 1{,}658$.

Визначимо фактичне значення t як відношення вибіркового коефіцієнта кореляції до середньої помилки вибіркового коефіцієнта кореляції:

$$t_\phi = \frac{z}{m} = \frac{1{,}658}{0{,}378} = 4{,}39.$$

Табличне значення t-критерію Стьюдента (дод. 2) при рівні значущості 0,05 і 8 ступенях свободи варіації ($n - 2 = 10 - 2$) дорівнює 2,307. Оскільки $t_\phi > t_{0{,}005}$ (4,39 > 2,307), то можна зробити висновок, що вибірковий коефіцієнт кореляції є вірогідним, істотним.

Обчислимо надійний інтервал, в якому при встановленому рівні значущості лежить коефіцієнт кореляції у генеральній сукупності:

$r = z \pm tm = 1{,}658 \pm 2{,}307 \cdot 0{,}378 = 1{,}658 \pm 0{,}872$, тобто від 0,786 до 2,530.

Користуючись таблицею значень z-критерію, визначимо межі надійного інтервалу для коефіцієнта кореляції у генеральній сукупності:

$$0{,}74 < r < 0{,}99.$$

Отже, з імовірністю помилитися в 5 випадках із 100 можна стверджувати, що тіснота зв'язку між урожайністю зернових культур та внесенням добрив у генеральній сукупності знаходиться в межах від 0,66 до 0,99.

Вибірковий коефіцієнт кореляції можна оцінити і простішими методами. Р.Фішером розроблено спеціальні таблиці, в яких наведено критичні значення коефіцієнта кореляції для різних рівнів значущості (0,05 і 0,01) та чисельності вибіркової сукупності. Щоб вибірковий коефіцієнт кореляції можна було вважати вірогідним, він повинен бути більшим за критичне значення, наведене в таблиці.

Орієнтовно можна вважати, що коли відношення коефіцієнта кореляції до його середньої помилки дорівнює 3 і більше, то коефіцієнт кореляції є вірогідним, тобто зв'язок між досліджуваними явищами є доведеним.

6.3 Криволінійна кореляція

Як зазначалося, при прямолінійній залежності спостерігається рівномірне збільшення (зменшення) результативної ознаки під впливом відповідної зміни факторної ознаки. У статистичній практиці трапляються та більш складні зв'язки, коли зі зміною аргументу змінюється не тільки функція, а й її приріст.

Нелінійні форми зв'язку різні. У статистичному аналізі найчастіше використовують параболічну та гіперболічну форми зв'язку.

Якщо криволінійна залежність має форму параболи другого порядку, зв'язок виражають таким рівнянням:

$$y_x = a_0 = a_1 x + a_2 x^2,$$

де y_x — теоретичні значення результативної ознаки; a_0, a_1, a_2 — параметри рівняння; x – значення факторної ознаки.

Параметри a_0, a_1 і a_2 визначають складанням та розв'язанням системи трьох рівнянь:

$$\sum y = na_0 + a_1 \sum x + a_2 \sum x^2;$$
$$\sum yx = a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3;$$
$$\sum yx^2 = a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4.$$

Щоб спростити розв'язання рівнянь, замість значень x введемо відхилення від середньої $(x - \bar{x})$. Тоді рівняння матимуть такий вигляд:

$$\sum y = na_0 + a_1 \sum (x - \bar{x}) + a_2 \sum (x - \bar{x})^2;$$
$$\sum y(x - \bar{x}) = a_0 \sum (x - \bar{x}) + a_1 \sum (x - \bar{x})^2 + a_2 \sum (x - \bar{x})^3;$$
$$\sum y(x - \bar{x})^2 = a_0 \sum (x - x)^2 + a_1 \sum (x - \bar{x})^3 + a_2 \sum (x - \bar{x})^4.$$

Оскільки $\sum (x - \bar{x})$ і $\sum (x - \bar{x})^3$ дорівнюють нулю, то після відповідних спрощень дістанемо

$$\sum y = na_0 + a_2 \sum (x - \bar{x})^2;$$
$$\sum y(x - \bar{x}) = a_1 \sum (x - \bar{x})^2;$$
$$\sum y(x - \bar{x})^2 = a_0 \sum (x - \bar{x})^2 + a^2 \sum (x - \bar{x})^4.$$

Порядок визначення параметрів рівняння розглянемо на такому прикладі (табл. 58).

Таблиця 58 *Вихідні та розрахункові дані для кореляційного аналізу залежності урожайності гречки від кількості опадів за вегетаційний період*

Опади, см y	Урожайність гречки, ц/га, y	$x - \bar{x}$	$(x - \bar{x})^2$	$y(x - \bar{x})^2$	$y(x - x)^2$	$(x - \bar{x})^4$	y_x
5	5,6	–15	225	–84	1260	50625	4,7
10	10,3	–10	100	–103	1030	10000	10,9
15	13,8	–5	25	–69	345	625	15,7
20	18,1	0	0	0	0	0	18,7
25	22,4	5	25	112	560	625	20,1
30	21,0	10	100	210	2100	10000	19,7
35	16,2	15	224	243	3645	50625	17,6
$\sum = 140$	107,4	—	700	309	8940	122500	107,4

Середня кількість опадів за вегетаційній період:

$$\bar{x}\frac{\sum x}{n} = \frac{140}{7} = 20.$$

Підставляємо дані табл. 58 у систему рівнянь:

$$107{,}4 = 7a_0 + 700a_2; \quad (1)$$
$$309 = 700a_1; \quad (2)$$
$$8940 = 700a_0 + 122500a_2 \quad (3)$$

Із другого рівняння визначимо, що параметр a_1 дорівнює 0,441. Перше і третє рівняння розділимо на коефіцієнт при a_0 (відповідно 7 і 700):

$$15{,}34 = a_0 + 100a_2; \quad (1)$$
$$12{,}77 = a_0 + 175a_2. \quad (2)$$

Від першого рівняння віднімаємо друге:

$$2{,}57 = -75a_2,$$
$$\text{звідки } a_2 = \frac{2{,}57}{-75} = -0{,}034.$$

Підставивши в одне з попередніх рівнянь значення параметра a_2, визначимо, що $a_0 = 18{,}74$.

Отже, рівняння параболи другого порядку, що характеризує залежність урожайності гречки від кількості опадів за вегетаційний період, матиме такий вигляд:

$$y_x = 18{,}74 + 0{,}441(x - \bar{x}) - 0{,}034(x - \bar{x})^2.$$

Підставивши в рівняння відповідні значення відхилень та їх квадрати, обчислимо теоретичні рівні урожайності (див. табл. 58).

На відміну від прямолінійної залежності коефіцієнти регресії криволінійного зв'язку не можна інтерпретувати однозначно, оскільки швидкість зміни результативної ознаки при різних значеннях факторної ознаки неоднакова. У нашому прикладі параметр $a_1 = 0{,}441$ показує приріст урожайності залежно від кількості опадів, що випали, а параметр $a_2 = -0{,}034$ характеризує зниження урожайності. Обчислені за рівнянням значення урожайності при кількості опадів 100 і 150 мм ($x = 10$ і $x = 15$) дорівнюють 10,9 і 15,7 ц/га, тобто приріст становить 4,8 ц. При тій самій різниці в кількості опадів, але при переході від 200 до 250 мм ($x = 20$ і $x = 25$) урожайність становить 18,7 і 20,1 ц/га, тобто приріст знизився до 1,4 ц.

Тісноту зв'язку при криволінійних формах залежності визначають за індексом кореляції (кореляційного відношення):

$$i = \sqrt{\frac{\sigma_{yx}^2}{\sigma_y^2}},$$

де σ_{yx}^2 — міжгрупова дисперсія; σ_y^2 — загальна дисперсія.

Індекс кореляції змінюється у межах від 0 до +1, тобто завжди є додатною величиною. Він показує, яку частку у загальному середньоквадратичному відхиленні результативної ознаки становить середньоквадратичне відхилення факторної ознаки.

У статистичній практиці найчастіше використовують такі робочі формули для визначення індексу кореляції:

$$i = \sqrt{\frac{\frac{\sum(y_x - \bar{y})^2}{n}}{\frac{\sum(y - \bar{y})^2}{n}}}; \quad i = \sqrt{\frac{\sum(y_x - \bar{y})^2}{\sum(y - \bar{y})^2}}; \quad i = \sqrt{\frac{\bar{y}_x^2 - \bar{y}^2}{\bar{y}^2 - \bar{y}^2}}.$$

Використовуючи дані табл. 58, визначимо індекс кореляції між кількістю опадів за вегетаційний період та урожайністю гречки (середня урожайність $\bar{y} = \bar{y}_x = 107{,}4 : 7 = 15{,}0$ ц/га).

$$i = \sqrt{\frac{\sum(y_x - \bar{y})^2}{\sum(y - \bar{y})^2}} = \sqrt{\frac{-10{,}6^2 + (-4{,}4^2) + 0{,}4^2 + 3{,}4^2 + 4{,}8^2 + 4{,}4^2 + 2{,}3^2}{-9{,}7^2 + (-5{,}0^2) + (-1{,}5^2) + 2{,}8^2 + 7{,}1^2 + 5{,}7^2 + 0{,}9^2}} =$$

$$= 0{,}948.$$

Обчислене кореляційне відношення показує, що між кількістю опадів за вегетаційний період та урожайністю гречки існує тісний зв'язок.

Індекс кореляції можна використати і для визначення тісноти зв'язу при прямолінійній залежності. У цьому разі абсолютна величина індексу кореляції збігається з лінійним коефіцієнтом кореляції. Якщо зв'язок криволінійний, то $i > r$. Математично встановлено, що коли різниця між індексом кореляції і коефіцієнтом кореляції не перевищує 0,1, то гіпотезу про прямолінійність зв'язку можна вважати доведеною.

Вірогідність (надійність) індексу кореляції визначають так само, як і коефіцієнта кореляції.

Іноді залежність між корелюючими величинами має гіперболічний характер. Її встановлюють за допомогою рівняння гіперболи

$$y_x = a_0 + \frac{a_1}{x},$$

де y_x — теоретичні значення результативної ознаки; a_0 і a_1 — параметри рівняння регресії; x — значення факторної ознаки.

Для того, щоб визначити параметри a_0 і a_1 способом найменших квадратів, розв'язують таку систему рівнянь з двома невідомими:

$$\sum y = na_0 + a_1 \sum \frac{1}{x};$$

$$\sum \frac{y}{x} = a_0 \sum \frac{1}{x} + a_1 \sum \frac{1}{x^2}.$$

Послідовність визначення параметрів гіперболічного рівняння розглянемо на такому прикладі (табл. 59).

Таблиця 59 *Кореляційний аналіз залежності витрат кормів на 1 ц молока від рівня продуктивності корів*

Річний надій молока від корови, ц	Середина інтервалу, x	Витрати кормів на 1 ц молока, ц корм. од., y	$\frac{1}{x}$	$\frac{1}{x^2}$	$\frac{y}{x}$	y_x
20 – 22	21	1,50	0,0476	0,00227	0,0714	1,48
22 – 24	23	1,42	0,0435	0,00189	0,0617	1,41
24 – 26	25	1,35	0,0400	0,00160	0,0540	1,34
26 – 28	27	1,29	0,0370	0,00137	0,0478	1,28
28 – 30	29	1,23	0,0345	0,00119	0,0424	1,23
30 – 32	31	1,18	0,0323	0,00104	0,0381	1,18
32 – 34	33	1,13	0,0303	0,00092	0,0342	1,15
34 – 36	35	1,10	0,0286	0,00082	0,0314	1,12
36 – 38	37	1,07	0,0270	0,00073	0,0289	1,09
38 – 40	39	1,05	0,0256	0,00066	0,0269	1,06
Разом	—	12,34	0,3464	0,01249	0,4368	12,34

Підставивши дані табл. 59 у рівняння, матимемо:

$$12{,}34 = 10a_0 + 0{,}3464a_1 \quad (1)$$
$$0{,}4368 = 0{,}3464a_0 + 0{,}01249a_1, \quad (2)$$

Помножимо перше рівняння на одницю, а друге — на 28,9 (10:0,3464):

$$12{,}34 = 10a_0 + 0{,}3464a_1; \quad (1)$$
$$12{,}62 = 10a_0 + 0{,}3610a_1. \quad (2)$$

Віднімемо від другого рівняння перше:

$$0{,}28 = 0{,}0146a_1.$$

Звідси $a_1 = \frac{0{,}28}{0{,}0146} = 19{,}2.$

Підставивши значення a1 в перше вихідне рівняння, обчислимо значення a_0:

$$12{,}34 = 10a_0 + 0{,}3464 \cdot 19{,}2;$$
$$12{,}34 = 10a_0 + 6{,}65;$$
$$10a_0 = 5{,}69;$$
$$a_0 = 0{,}57.$$

Отже, рівняння гіперболи, яке характеризує залежність витрат кормів на 1 ц молока від рівня продуктивності корів матиме такий вигляд:

$$y_x = 0{,}57 + 19{,}2\frac{1}{x}.$$

Підставляючи у це рівняння значення факторної ознаки x, дістанемо теоретичні рівні витрат кормів на виробництво 1 ц молока (див. табл. 59).

Індекс кореляції та його вірогідність при гіперболічній формі зв'язку визначають за тими самими формулами, що й при параболічній залежності.

Досліджуючи кореляційні зв'язки, слід уникати необґрунтованого ускладнення застосовуваних криволінійних рівнянь. Ускладнення рівнянь зв'язку, якщо воно не ґрунтується на достатньому логічному та економічному змісті досліджуваних явищ, утруднює інтерпретацію показників регресії та позбавляє аналіз цілеспрямованості.

6.4 Множинна кореляція

Визначення та кількісна оцінка взаємозв'язку між двома статистичними ознаками за допомогою парної кореляції є дійовим засобом статистичного аналізу. Проте соціально-економічні процеси та явища формуються під впливом не одного, а багатьох факторів. Наприклад, на урожайність сільськогосподарських культур впливають метеорологічні умови, кількість унесених добрив, сорт, строки сівби тощо. Продуктивність тварин залежить від рівня та якості годівлі, породи, способів утримання тварин, процесів відтворення стада тощо.

Кореляцію, за допомогою якої вивчається вплив на результативну ознаку двох та більше взаємозв'язаних факторних ознак, називають множинною. При вивченні множинної кореляції можна застосовувати як прямолінійні, так і криволінійні рівняння регресії.

Багатофакторні регресійні моделі дають змогу оцінювати вплив на досліджувану результативну ознаку кожного окремого із включених у рівняння факторів при фіксованому значенні (на середньому рівні) інших факторів. При цьому важливою умовою множинної кореляції є відсутність функціонального зв'язку між факторами.

Важливе значення при множинній кореляції має вибір форми зв'язку та відповідного математичного рівняння множинної регресії. Вибір типу функції має грунтуватися на теоретичному аналізі досліджуваного явища або на досвіді попередніх аналогічних досліджень. Ураховуючи, що будь-яку функцію багатьох змінних можна звести до лінійного типу логарифмуванням, рівняння множинної регресії частіше будують у лінійній формі.

Формула лінійного рівняння множинної регресії має такий вигляд:

$$y_x = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n,$$

де y_x — теоретичне значення результативної ознаки; a_0, a_1, a_2 ...a_n параметри рівняння; x_1, x_2 ...x_n — факторні ознаки.

Окремі коефіцієнти регресії цього рівняння характеризують вплив відповідного фактора на результативний показник при фіксованому (елімінованому) значенні інших факторів. Вони показують, наскільки змінюється результативний показник при зміні відповідного фактора на одиницю. Вільний член рівняння (a_0) не має економічного змісту та не інтерпретується.

Параметри рівняння множинної регресії обчислюють способом найменших квадратів розв'язанням системи рівнянь:

$$\sum y = na_0 + a_1\sum x_1 + a_2\sum x_2 + \ldots + a_n\sum x_n;$$
$$\sum yx_1 = a_0\sum x_1 + a_1\sum x_1^2 + a_2\sum x_1x_2 + \ldots a_n\sum x_1x_n;$$
$$\sum yx_2 = a_0\sum x_2 + a_1\sum x_1x_2 + a_2\sum x_2^2 + \ldots a_n\sum x_2x_n^2;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\sum yx_n = a_0\sum x_n + a_1\sum x_1x_n + a_2\sum x_2x_n + \ldots + a_n\sum x_n^2.$$

Порядок обчислення рівняння множинної лінійної регресії розглянемо на прикладі даних про залежність урожайності зернових культур від якості грунтів та унесених мінеральних добрив (табл. 60).

Перенесемо підсумкові дані табл. 60 у систему рівнянь:

$$398 = 10a_0 + 684a_1 + 17{,}3a_2;$$
$$27287 = 684a_0 + 46892a_1 + 1188{,}5a_2;$$
$$693{,}4 = 17{,}3a_0 + 1188{,}5a_1 + 31{,}09a_2.$$

Розділимо кожне рівняння на коефіцієнт при a_0:

$$39{,}80 = a_0 + 68{,}4a_1 + 1{,}730a_2; \quad (1)$$
$$39{,}89 = a_0 + 68{,}56a_1 + 1{,}738a_2; \quad (2)$$
$$40{,}08 = a_0 + 68{,}70a_1 + 1{,}797a_2. \quad (3)$$

Таблиця 60 *Вихідні та розрахункові дані для множинної кореляції*

Номер господарства	Якість грунтів, балів x_1	Унесення мінеральних добрив на 1 га посівів, ц д.р., x_2	Урожайність зернових культур, ц/га y	x_1^2	x_2^2	x_1x_2	yx_1	yx_2	y^2
1	70	1,7	41	4900	2,89	119,0	2870	69,7	1681
2	67	2,0	38	4489	4,00	134,0	2546	76,0	1444
3	69	1,7	42	4761	2,89	117,3	2898	71,4	1764
4	68	1,5	39	4624	2,25	102,0	2652	58,5	1521
5	68	1,2	38	4624	1,44	81,6	2584	45,6	1444
6	65	2,0	37	4225	4,00	130,0	2405	74,0	1369
7	70	2,2	45	4900	4,84	154,0	3150	99,0	2025
8	75	2,2	42	5625	4,84	165,0	3150	92,4	1764
9	62	1,3	36	3844	1,69	80,6	2232	46,8	1296
10	70	1,5	40	4900	2,25	105,0	2800	60,0	1600
Разом	684	17,3	398	46892	31,09	1188,5	27287	693,4	15908
У середньому	68,4	1,73	39,8	4689,2	3,109	118,85	2728,7	69,34	1590,8

Від другого рівняння віднімемо перше і третє:

$$0,09 = 0,16a_1 + 0,008a_2;$$
$$-0,19 = -0,14a_1 - 0,059a_2.$$

Розділимо одержані рівняння на коефіцієнт при a_1:

$$0,562 = a_1 + 0,050a_2; \quad (1)$$
$$1,357 = a_1 + 0,421a_2. \quad (2)$$

Від другого рівняння віднімемо перше:

$$0,795 = 0,371a_2.$$

Звідси

$$a_2 = \frac{0,795}{0,371} = 2,14.$$

Підставивши значення параметра a_2 в одне з наведених вище рівнянь, визначимо параметр a_1:

$$0,562 = a_1 + 0,050 \cdot 2,14;$$
$$0,562 = a_1 + 0,107;$$
$$a_1 = 0,562 - 0,107 = 0,45.$$

Значення параметрів $a_1 = 0,45$ і $a_2 = 2,14$ підставимо в одне з проміжних рівнянь і обчислимо значення a_0:

$$39,80 = a_0 + 68,4 \cdot 0,45 + 1,73 \cdot 2,14;$$
$$39,80 = a0 + 30,78 + 3,70;$$
$$a_0 = 39,80 - 30,78 - 3,70;$$
$$a_0 = 5,32.$$

Отже, рівняння множинної регресії, яке характеризує залежність урожайності зернових культур від якості грунтів та унесених мінеральних добрив, матиме такий вигляд:

$$y_{x_{1,2}} = 5{,}32 + 0{,}45x_1 + 2{,}14x_2.$$

Коефіцієнти регресії показують, наскільки зміниться урожайність при зміні відповідного фактора на одиницю при умові, що другий фактор, включений у рівняння, лежить на середньому рівні. Так, $a_1 = 0{,}45$ показує, що при середній дозі унесених мінеральних добрив поліпшення якості грунтів на один бал сприяє зростанню врожайності зернових культур на 0,45 ц/га. Збільшення дози унесених мінеральних добрив на 1 ц при середній якості грунтів забезпечує приріст урожайності на 2,14 ц/га.

Підставивши у рівняння множинної регресії фактичні значення змінних x по кожному господарству, визначимо теоретичні рівні урожайності зернових культур. Так, для першого господарства теоретична урожайність

$$y_x = 5{,}32 + 0{,}45 \cdot 70 + 2{,}14 \cdot 1{,}7 = 40{,}5 \text{ ц/га}.$$

Обчислені за рівнянням рівні врожайності та їх квадрати наведено в табл. 61.

Показниками тісноти зв'язку при множинній кореляції є парні, часткові та множинні (сукупні) коефіцієнти кореляції та множинний коефіцієнт детермінації.

Парні коефіцієнти кореляції використовують для вимірювання тісноти зв'язку між двома досліджуваними ознаками без урахування їх взаємодії з іншими ознаками, включеними в кореляційну модель. Методика розрахунку цих коефіцієнтів та їх інтерпретація така сама, як і методика розрахунку лінійного коефіцієнта парної кореляції при однофакторному зв'язку.

Таблиця 61 *Теоретичні рівні урожайності зернових культур та їх квадрати*

Показник / Номер господарства	Урожайність зернових культур, ц/га	Квадрат урожайності зернових культур
1	40,5	1640,25
2	39,8	1584,04
3	40,0	1600,00
4	39,1	1528,81
5	38,5	1482,25
6	38,8	1505,44
7	41,5	1722,25
8	43,8	1918,44
9	36,0	1296,00
10	40,0	1600,00
Разом	398,0	15877,48
У середньому	39,8	1587,7

У нашому прикладі коефіцієнти парної кореляції дорівнюватимуть:

- між урожайністю та якістю грунтів:

$$r_{yx_1} = \frac{\overline{x_1 y} - \bar{x}_1 \cdot \bar{y}}{\sqrt{\left(\overline{x_1^2} - \bar{x}_1^2\right)\left(\overline{y^2} - \bar{y}^2\right)}} = \frac{2728{,}7 - 68{,}4 \cdot 39{,}8}{\sqrt{\left(4689{,}2 - 68{,}4^2\right) \cdot \left(1590{,}8 - 39{,}8^2\right)}} = 0{,}752;$$

- між урожайністю та мінеральними добривами:

$$r_{yx_2} = \frac{\overline{x_2 y} - \bar{x}_2 \cdot \bar{y}}{\sqrt{\left(\overline{x_2^2} - \bar{x}_2^2\right)\left(\overline{y^2} - \bar{y}^2\right)}} = \frac{69{,}34 - 1{,}73 \cdot 39{,}8}{\sqrt{\left(3{,}109 - 1{,}73^2\right) \cdot \left(1590{,}8 - 39{,}8^2\right)}} = 0{,}549;$$

- між якістю грунтів та мінеральними добривами:

$$r_{x_1x_2} = \frac{\overline{x_1 x_2} - \bar{x}_1 \cdot \bar{x}_2}{\sqrt{\left(\overline{x_1^2} - \bar{x}_1^2\right)\left(\overline{x_2^2} - \bar{x}_2^2\right)}} = \frac{118{,}85 - 68{,}4 \cdot 1{,}73}{\sqrt{\left(4689{,}2 - 68{,}4^2\right) \cdot \left(3{,}109 - 1{,}73^2\right)}} = 0{,}466;$$

Обчислені парні коефіцієнти кореляції показують, що урожайність зернових культур перебуває в тісному зв'язку як із якістю грунтів $\left(r_{yx_1} = 0{,}752\right)$, так із унесенням мінеральних добрив $(r_{yx_2} = 0{,}549)$. Існує також невелика залежність та між факторними ознаками: якістю грунтів та мінеральними добривами $\left(r_{x_1x_2} = 0{,}466\right)$.

Кореляційний зв'язок між факторами в рівнянні множинної регресії називають **колінеарністю** або **мультиколінеарністю**. Мультиколінеарність ускладнює вивчення впливу окремих факторів на результативну ознаку, оскільки взаємодія колінеарних факторів у моделі подвоюватиметься і спотворюватиме результати. Чим вища колінеарність, тим менш надійними будуть показники впливу окремих факторів.

Статистикою ще не знайдені надійні способи вимірювання мультиколінеарності. Допустимою колінеарністю для практичних цілей, що не спотворює результати дослідження, вважають таку, при якій парні коефіцієнти кореляції між факторними та результативними ознаками більші за коефіцієнт кореляції між супутніми факторами:

$$r_{yx_1} > r_{x_1x_2}; \qquad r_{yx_2} > r_{x_1x_2};$$

Часткові коефіцієнти кореляції характеризують тісноту зв'язку результативної ознаки з однією факторною ознакою при умові, що інші факторні ознаки перебувають на постійному рівні. Парний коефіцієнт кореляції між результативною та факторною ознаками, як правило, відрізняється від відповідного часткового коефіцієнта.

Частковий коефіцієнт кореляції між ознаками y і x_1 без урахування впливу ознаки x_2 визначають за формулою

$$r_{yx(x_2)} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}},$$

де r — парні коефіцієнти кореляції між відповідними ознаками.

У нашому прикладі частковий коефіцієнт кореляції між урожайністю зернових культур та якістю грунтів

$$r_{yx_1} = \frac{0{,}752 - 0{,}549 \cdot 0{,}466}{\sqrt{\left(1 - 0{,}549^2\right) \cdot \left(1 - 0{,}466^2\right)}} = 0{,}740.$$

Частковий коефіцієнт кореляції між урожайністю та мінеральними добривами

$$r_{yx_2} = \frac{0{,}549 - 0{,}752 \cdot 0{,}466}{\sqrt{\left(1 - 0{,}752^2\right) \cdot \left(1 - 0{,}466^2\right)}} = 0{,}584.$$

Додатні знаки перед частковими коефіцієнтами кореляції свідчать про пряму залежність між досліджуваними ознаками. Визначення часткових коефіцієнтів дає змогу глибше вивчити зв'язок між явищами і, зокрема, виявити вплив факторної ознаки на зміну результативної ознаки.

Коефіцієнт множинної (сукупної) детермінації показує, яка частка варіації досліджуваного результативного показника

зумовлена впливом факторів, включених у рівняння множинної регресії. Він може мати значення від 0 до +1. Чим ближчий коефіцієнт множинної детермінації до одиниці, тим більше варіація результативного показника характеризується впливом відібраних факторів. Коефіцієнт множинної детермінації визначають за такою формулою:

$$R^2 = \frac{\sigma^2_{обч}}{\sigma^2_з},$$

де $\sigma^2_{обч}$ — дисперсія результативного показника, обчислена за рівнянням множинної регресії; $\sigma^2_з$ — загальна дисперсія результативного показника.

У нашому прикладі

$$\sigma^2_{обч} = \overline{y^2_x} - \overline{y}^2 = 1587{,}7 - 39{,}8^2 = 3{,}66;$$
$$\sigma^2_з = \overline{y^2} - \overline{y}^2 = 1590{,}8 - 39{,}8^2 = 6{,}76.$$

Отже,

$$R^2 = \frac{3.66}{6.76} = 0{,}541, \text{ або } 54{,}1\ \%.$$

Коефіцієнт множинної детермінації показує, що 54.1 % варіювання урожайності зернових культур у досліджуваних господарствах зумовлене включеними у кореляційну модель факторами: якістю грунтів та унесеними мінеральними добривами. Решта коливання урожайності (45,9 %) зумовлена іншими факторами (метеорологічними умовами, сортом, строками сівби тощо).

Основним показником тісноти зв'язку при множинній кореляції є **коефіцієнт множинної (сукупної) кореляції**, який може мати значення від 0 до +1. Формула для його обчислення має такий вигляд:

$$R = \sqrt{\frac{\sigma^2_{обч}}{\sigma^2_з}} = \sqrt{0{,}541} = 0{,}736.$$

При лінійному двофакторному зв'язку коефіцієнт множинної кореляції можна визначити за такою формулою:

$$R = \sqrt{\frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r^2_{x_1x_2}}},$$

де r — лінійні парні коефіцієнти кореляції.

У нашому прикладі

$$R = \sqrt{\frac{0{,}752^2 + 0{,}549^2 - 2 \cdot 0{,}752 \cdot 0{,}549 \cdot 0{,}466}{1 - 0{,}466^2}} = 0{,}785.$$

Деяка розбіжність між коефіцієнтами множинної кореляції, обчисленими різними способами, свідчить про те, що у нашому прикладі зв'язок між корелюючими величинами не є абсолютно прямолінійним, а має елементи криволінійності. Правильним буде коефіцієнт, обчислений як корінь квадратний з коефіцієнта множинної детермінації.

Середню помилку вибіркового коефіцієнта множинної кореляції визначають за формулою

$$m_R = \frac{1 - R^2}{\sqrt{n - m - 1}},$$

де n — кількість спостережнь; m — кількість факторів.

У нашому прикладі середня помилка коефіцієнта множинної кореляції:

$$m_R = \frac{1-0{,}736^2}{\sqrt{10-2-1}} = \frac{0{,}459}{\sqrt{2{,}65}} = 0{,}173.$$

Фактичне значення t-критерію:

$$t_\phi = \frac{R}{m_R} = \frac{0{,}736}{0{,}173} = 4{,}254.$$

Табличне значення t-критерію при рівні ймовірності 0,95 і 7 ступенях свободи дорівнює 2,3646 (дод. 2). Оскільки фактичне значення t значно перевищує табличне, то можна зробити висновок про вірогідність коефіцієнта множинної кореляції.

Важливими показниками кореляційного аналізу є **коефіцієнти еластичності** і **β-коефіцієнти**. Потреба в їх застосуванні зумовлена тим, що коефіцієнти регресії, маючи різні фізичний зміст і одиниці вимірювання, не дають чіткого уявлення про те, які фактори мають найбільший вплив на результативну ознаку, тобто коефіцієнти регресії не можна безпосередньо порівнювати між собою.

Коефіцієнти еластичності (E) показують, на скільки процентів змінюється результативна ознака при зміні факторної ознаки на 1 %. Їх обчислюють за такою формулою:

$$E = a_1 \frac{\bar{x}_i}{\bar{y}},$$

де a_1 коефіцієнт регресії при i-му факторі; $\bar{x}_i$ — середнє значення і-го фактора; $\bar{y}$ — середнє значення результативної ознаки.

У нашому прикладі (див. табл. 60) коефіцієнти еластичності дорівнюватимуть:

- між урожайністю зернових культур та якістю грунтів

$$E_1 = \frac{0{,}45 \cdot 68{,}4}{39{,}8} = 0{,}77;$$

- між урожайністю та унесеними мінеральними добривами

$$E_2 = \frac{2{,}14 \cdot 1{,}73}{39{,}8} = 0{,}09.$$

Обчислені коефіцієнти показують, що при поліпшенні якості грунтів на 1 % урожайність зернових культур зростає на 0,77 %, а при збільшенні дози унесених мінеральних добрив на 1 % урожайність зростає на 0,09 %.

β-коефіцієнти показують, на скільки середніх квадратичних відхилень змінюється результативний показник при зміні відповідного фактора на одне значення середньоквадратичного відхилення. Вони характеризують вплив окремих факторів на результативну ознаку. Їх визначають за формулою

$$\beta_i = a_1 \frac{\sigma_{x_1}}{\sigma_y},$$

де a_1 — коефіцієнт регресії при і-му факторі; σ_{x_i} — середнє квадратичне відхилення i-го фактора; σ_y — середнє квадратичне відхилення результативного показника.

У нашому прикладі (див. табл. 60) середнє квадратичне відхилення:

- урожайності $\sigma_y = \sqrt{\bar{y}^2 - \bar{y}^2} = \sqrt{1590.8 - 39.8^2} = 2.60$;
- якості грунтів $\sigma_{x_1} = \sqrt{\bar{x}_1^2 - \bar{x}^2} = \sqrt{4689{,}2 - 68{,}4^2} = 3{,}26$;
- мінеральних добрив $\sigma_{x_2} = \sqrt{\bar{x}_2^2 - \bar{x}^2} = \sqrt{3{,}109 - 1{,}73^2} = 0{,}34$.

β-коефіцієнти:

- між урожайністю та якістю грунтів

$$\beta_1 = \frac{0{,}45 \cdot 3{,}26}{2{,}60} = 0{,}56;$$

- між урожайністю та мінеральними добривами

$$\beta_2 = \frac{2{,}14 \cdot 0{,}34}{2{,}60} = 0{,}28;$$

Обчислені β-коефіцієнти показують, що при поліпшенні якості грунтів на одне середньоквадратичне відхилення врожайність зернових культур зростає на 0,56 середньоквадратичного відхилення, а при відповідному збільшенні дози унесених мінеральних добрив — на 0,28 середньоквадратичного відхилення. Отже, найбільші резерви досліджуваної групи господарств щодо підвищення врожайності зернових культур пов'язані з поліпшенням якості грунтів.

β-коефіцієнти використовують для розкладання загальної варіації результативного показника на включені у кореляційну модель фактори. Для цього визначають парні коефіцієнти детермінації як добуток парних коефіцієнтів кореляції на β-коефіцієнти відповідних факторів (табл. 62).

Таблиця 62 *Розкладання загальної варіації урожайності на фактори*

Фактор	r_{yx_i}	β_{x_i}	Коефіцієнт	Процент
Якість грунтів	0,752	0,56	0,421	42,1
Мінеральні добрива	0,549	0,28	0,154	15,4

Дані табл. 62 підтверджують раніше зроблений висновок про те, що основним фактором, що впливає на урожайність зернових культур, є якість грунтів: 42,1 % загального коливання урожайності зумовлене цим фактором.

6.5 Непараметричні критерії кореляційних зв'язків

У попередніх параграфах цього розділу розглянуто методи вимірювання зв'язку між результативними та факторними ознаками для сукупностей, які мають нормальний або близький до нормального розподіл. Якщо ж характер розподілу досліджуваної сукупності невідомий, тісноту кореляційного зв'язку визначають за допомогою **непараметричних методів**. Особливістю цих методів є те, що коефіцієнт кореляції між ознаками визначають не за кількісними значеннями варіантів ознак, а за допомогою порівняння їх рангів. Під рангом розуміють порядковий номер відповідної одиниці сукупності у ранжированому ряду. Чим менша розбіжність між порядковими номерами порівнюваних ознак, тим тісніший зв'язок між ними.

Непараметричні критерії використовують для вимірювання тісноти зв'язку якісних ознак, які не мають кількісного виразу, але їх можна ранжирувати. Із непараметричних критеріїв тісноти зв'язку найчастіше застосовують коефіцієнт кореляції рангів та коефіцієнт Фехнера.

Коефіцієнт кореляції рангів обчислюють за формулою Спірмена

$$r_p = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)},$$

де d — різниця між рангами елементів сукупності за першою та другою ознаками; n — кількість спостережень.

Коефіцієнт кореляції рангів може мати значення від 0 до +1 і від 0 до –1. Якщо обидва ряди рангів повністю збігаються, то $\sum d^2 = 0$ і коефіцієнт кореляції дорівнює +1. При повному зворотному зв'язку, коли ранги розташовані у зворотному порядку, коефіцієнт кореляції дорівнює –1.

Обчислення коефіцієнта кореляції рангів розглянемо на прикладі рангів 10 абітурієнтів згідно з їх середніми оцінками за атестатом зрілості та на вступних іспитах у ВНЗ (табл. 63).

Таблиця 63 *Вихідні та розрахункові дані для обчислення коефіцієнта рангової кореляції*

Абітурієнт	Ранг		Різниця рангів	Квадрат різниці
	атестат зрілості	вступні іспити		
	R_x	R_y	d	d^2
1	4	3	–1	1
2	2	1	–1	1
3	5	6	+1	1
4	1	2	+1	1
5	7	7	0	0
6	9	8	–1	1
7	3	4	+1	1
8	10	9	–1	1
9	8	10	+2	4
10	6	5	–1	1
Разом	55	55	—	12

Підставивши дані табл. 63 у формулу Спірмена, дістанемо:

$$r_p = 1 - \frac{6 \cdot 12}{10 \cdot \left(100 - 1\right)} = 0{,}927.$$

Обчислений коефіцієнт кореляції рангів свідчить про прямий тісний зв'язок між успішністю абітурієнтів у школі та на вступних іспитах у ВНЗ.

Для того щоб перевірити істотність зв'язку, порівняємо обчислений коефіцієнт кореляції з його критичним значенням. При рівні істотності 0,05 і 8 ступенях свободи варіації (10 – 2) критичне значення коефіцієнта кореляції дорівнює 0,632. Оскільки фактичне значення коефіцієнта кореляції більше за критичне, то можна вважати, що між досліджуваними ознаками існує тісний зв'язок.

Коефіцієнт Фехнера застосовують для оцінки тісноти зв'язку порівнянням знаків відхилень варіантів факторної та результативної ознак від середніх. Його обчислюють за такою формулою:

$$r_з = \frac{\sum З - \sum Н}{\sum З + \sum Н},$$

де $\sum З$ — кількість збігів знаків; $\sum Н$ — кількість незбігів знаків у відхиленнях від середніх.

Послідовність розрахунку коефіцієнта кореляції розглянемо на прикладі даних про результати складання абітурієнтами вступних іспитів та їх успішності за першу екзаменаційну сесію у ВНЗі (табл. 64).

Підставивши дані табл. 64 у формулу коефіцієнта Фехнера, матимемо:

$$r_з = \frac{8-2}{8+2} = 0{,}6.$$

Обчислений коефіцієнт свідчить про помітний прямий зв'язок між результатами двох іспитів.

Коефіцієнт Фехнера змінюється від 0 до +1 та від 0 до –1. Якщо він дорівнює +1, то це означає, що повністю узгоджена пряма мінливість, при 0 — мінливість повністю не узгоджується, а при –1 — повна зворотна узгодженість мінливості. Слід мати на увазі, що коефіцієнт Фехнера враховує тільки напрям варіації та не враховує її величину, а тому оцінка тісноти зв'язку є приблизною.

Таблиця 64 *Розрахунок коефіцієнта Фехнера*

Студент	Середній бал вступних іспитів	Середній бал успішності за першу екзаменаційну сесію	Знаки відхилення за ознаками		Збіг знаків	
			факторна	Результативна	так	ні
	x	y	$x - \bar{x}$	$y - \bar{y}$	З	Н
1	4,0	3,0	–	–	З	
2	4,5	4,5	+	+	З	
3	5,0	5,0	+	+	З	
4	3,5	4,0	–	–	З	
5	4,0	3,5	–	–	З	
6	4,5	4,5	+	+	З	
7	3,5	3,5	–	–	З	
8	4,0	4,5	–	+		Н
9	4,5	5,0	+	+	З	
10	4,5	3,5	+	–		Н
Разом	42,0	41,0	–	–	8	2

6.6 Кореляція якісних ознак

Для вимірювання тісноти зв'язку якісних варіюючих ознак використовують коефіцієнт взаємного узгодження та коефіцієнт асоціації.

Коефіцієнт взаємного узгодження, запропонований О. О. Чупровим і доповнений Крамером, застосовують для вимірювання тісноти зв'язку між двома атрибутивними ознаками, коли їх варіювання утворює кілька (більше двох) груп. Наприклад, потрібно визначити тісноту зв'язку між результатами захисту дипломних робіт студентами та їх участю в роботі наукових студентських гуртків. Для цього проведено обстеження 150 студентів-випускників, результати якого зведені у комбінаційній таблиці, що дістала назву таблиці співзалежності (табл. 65).

Таблиця 65 *Розподіл 150 студентів-випускників за результатами захисту дипломних робіт та участі в наукових гуртках*

Результати захисту дипломних робіт	У наукових гуртках		Разом
	брали участь	не брали участі	
Відмінно	40	5	45
Добре	50	25	75
Задовільно	10	20	30
Разом	100	50	150

Коефіцієнт взаємного узгодження за формулою Чупрова

$$C = \sqrt{\frac{x^2}{n\sqrt{(m_1 - 1)(m_2 - 1)}}},$$

де x^2 — показник співзалежності; n — кількість елементів сукупності; m1 — кількість груп за першою ознакою; m_2 — кількість груп за другою ознакою.

При функціональному зв'язку коефіцієнт взаємного узгодження дорівнює одиниці тільки за умови рівності кількості груп, тобто коли $m_1 = m_2$. Якщо кількість груп кожної ознаки $m_1 \neq m_2$, то цей коефіцієнт менший за одиницю навіть при функціональному зв'язку. У такому разі коефіцієнт взаємного узгодження обчислюють за формулою Крамера:

$$C = \sqrt{\frac{x^2}{n(m-1)}},$$

де $m = \min (m, m_2)$.

Таблиця 66 *Розрахунок коефіцієнта взаємного узгодження*

Група студентів за результатами захисту дипломних робіт	Група студентів за участю в наукових гуртках		Разом
	1	2	
1	40 30	5 15	45
2	50 50	25 25	75
3	10 20	20 10	30
Разом	100	50	150

Якщо $m_1 = m_2$, то значення коефіцієнтів, обчислених за формулами Чупрова і Крамера збігаються. Оскільки у нашому прикладі кількість груп за досліджуваними ознаками різна, то коефіцієнт взаємного узгодження потрібно визначати за формулою Крамера порів-

нянням фактичних частот з теоретичними. Теоретичні частоти для першого рядка становитимуть:

$$f_1' = \frac{45 \cdot 100}{150} = 30. \quad f_2' = \frac{45 \cdot 100}{150} = 15.$$

Обчислені таким способом теоретичні частоти наведено поряд з фактичними в табл. 66.

$$x^2 = \sum \frac{(f - f')}{f'} = \frac{(40-30)^2}{30} + \frac{(5-15)}{15} + \frac{(50-50)^2}{50} +$$
$$+ \frac{(25-25)^2}{25} + \frac{(10-20)^3}{20} + \frac{(20-10)^2}{10} = 20.$$

Коефіцієнт взаємного узгодження за формулою Крамера

$$C = \sqrt{\frac{x^2}{n(m-1)}} = \sqrt{\frac{20}{150 \cdot (2-1)}} = 0{,}365.$$

Цей показник змінюється від 0 до 1. Проте при значенні більш як 0,3 можна робити висновок про тісний зв'язок між варіацією досліджуваних ознак. У нашому прикладі коефіцієнт взаємного узгодження показує тісний зв'язок між участю студентів у роботі наукових гуртків та результатами захисту ними дипломних робіт.

Істотність зв'язку перевіримо за допомогою критерію χ^2. Кількість ступенів свободи варіації у нашому прикладі дорівнює

$$(m_1 - 1) \cdot (m_2 - 1) = (3 - 1) \cdot (2 - 1) = 2.$$

При рівні значущості 0,05 і 2 ступенях свободи варіації критичне значення χ^2 становить 5,991. Оскільки фактичне значення χ^2 (20) значно перевищує критичне, то можна вважати, що істотність зв'язку доведено.

Для того, щоб визначити тісноту зв'язку між двома альтернативними ознаками ($m_1 = m_2 = 2$), застосовують коефіцієнт асоціації. Для цього вихідні дані зводять у комбінаційну таблицю співзалежності. Якщо дані кожної з чотирьох клітин позначити через *a, b, c* і *d*, то формула для обчислення коефіцієнта асоціації матиме такий вигляд:

$$K_a = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Коефіцієнт асоціації уперше запропонований англійським статистиком К.Пірсоном і в статистичній літературі відомий ще під назвою **коефіцієнта контингенції**.

Обчислення коефіцієнта асоціації розглянемо на такому прикладі (табл. 67).

Таблиця 67 *Розподіл 200 студентів-заочників за результатами складання іспиту із спеціального предмета та характером роботи*

Результати іспиту	Характер роботи		Разом
	за спеціальністю	не за спеціальністю	
Склали	140	30	170
Не склали	10	20	30
Разом	150	50	200

Підставивши дані табл. 67 у формулу, матимемо:

$$K_a = \frac{140 \cdot 20 - 30 \cdot 10}{\sqrt{170 \cdot 30 \cdot 150 \cdot 50}} = 0{,}404.$$

Коефіцієнт асоціації змінюється від 0 до +1 при прямій залежності і від 0 до –1 – при звоторній залежності. Якщо він більше 0,3, то це свідчить про істотний зв'язок між досліджуваними ознаками.

Щоб перевірити істотність зв'язку, використовують критерій χ^2, який функціонально пов'язаний з коефіцієнтом асоціації:

$$\chi^2 = K_a^2 n = 0{,}404^2 \cdot 200 = 32{,}64.$$

При рівні значущості 0,05 і 1 ступені свободи варіації $[(m_1 - 1)(m_2 - 1] = [(2 - 1)(2 - 1)]$ критичне значення дорівнює 3,84. Оскільки фактичне значення χ^2 більше за табличне, то можна зробити висновок про істотність зв'язку між результатами складання іспиту із спеціального предмета студентами-заочниками та характером їх роботи.

6.7 Особливості кореляційного аналізу в рядах динаміки

Аналіз кореляційних зв'язків у динамічних рядах має певні методологічні особливості, зумовлені взаємозалежністю рівнів, їх автокореляцією. Під **автокореляцією** розуміють кореляцію значень y_t і $y_t + h$, де t — потоковий момент часу, h — період часу. Автокореляція впливає на незалежність спостережень та спричинює викривлення результатів аналізу.

Існують різні способи усунення автокореляції. Найпростішим з них є спосіб так званої **кореляції по різницях**, запропонований Гукером. Суть цього способу полягає в заміні первинних рівнів взаємопов'язаних рядів динаміки y_t і x_t абсолютними приростами, тобто різницями. Так, різниці першого порядку $\Delta y = y_t - y_{t-1}$ і $\Delta x = x_t - x_{t-1}$ усувають лінійний тренд, а рівняння регресії матиме такий вигляд:

$$\Delta y = a_0 + a_1 \Delta x,$$

де a_0 — початок відліку; a_1 — коефіцієнт регресії (коефіцієнт пропорційності).

Якщо тенденція в рядах динаміки має нелінійний характер (результативна ознака змінюється зі сповільненням або ирискоренням порівняно із зміною факторної ознаки), слід застосовувати **спосіб відхилень від тенденції**. Суть його полягає в тому, що замість первинних рівнів y_t і x_t використовують їх відхилення від теоретичних рівнів, обчислених за рівняннями відповідних ліиій.

Одним з методів усунення автокореляції є введения у рівняння регресії змінної величини t, яка виконує роль фактора часу. Побудову такого рівняння розглянемо на прикладі двох взаємопов'язаних рядів динаміки, що характеризують зміну енергозабезпеченості господарства й урожайності зернових культур (табл. 68).

Таблиця 68 *Вихідні та розрахункові дані для обчислення коефіцієнтів регресії в рядах динаміки*

Пор. № року	Припадає енергоресурсів на 100 га посівної площі, к.с. x	Урожайність зерно-вих культур, ц/га y	t	xy	x^2	t^2	xt	yt	y_t
1	412	32	1	13184	169744	1	412	32	33,5
2	426	35	2	14910	181476	4	852	70	34,8
3	434	34	3	14756	188356	9	1302	102	35,7
4	440	36	4	15840	193600	16	1760	144	36,4
5	445	39	5	17355	198025	25	2225	195	37,1
6	454	35	6	15890	206116	36	2724	210	38,0
7	460	40	7	18400	211600	49	3220	280	38,8
8	455	39	8	17745	207025	64	3640	312	38,7
9	455	42	9	19110	207025	81	4095	378	39,0
Разом	3981	332	45	147390	1762967	285	20230	1723	332

Дані табл. 68 свідчать про те, що забезпеченість господарства енергетичними потужностями, крім останніх двох років, систематично підвищувалася; динаміка урожайності зернових кул ьтур була більш складною, з помітною тенденцією до її зростання. Отже, підвищення уро-

жайності зумовлене не тільки підвищенням енергозабезпеченості, а й іншими факторами, зокрема збільшенням доз унесених добрив, поліпшенням сортового складу, розширенням площ зрошуваних земель, кліматичними умовами тощо. Зв'язок між цими показниками описується лінійною функцією типу

$$yt = a_0 + a_1 x + a_2 t,$$

де yt — теоретичний рівень урожайності; a_0 — початок відліку; a_1 — коефіцієнт регресії, який характеризує середній приріст урожайності на одиницю приросту енергозабезпеченості; a_2 — коефіцієнт регресії, який характеризує щорічний приріст урожайності під впливом зміни комплексу факторів, крім x; x — рівень енергозабезпеченості, к.с.; t — показник часу.

Невідомі параметри рівняння регресії визначають методом найменших квадратів за допомогою складання та розв'язання системи нормальних рівнянь:

$$\sum y = na_0 + a_1 \sum x + a_2 \sum t;$$
$$\sum yx = a_0 \sum x + a_1 \sum x^2 + a_2 \sum xt;$$
$$\sum yt = a_0 \sum t + a_1 \sum xt + a_2 \sum t^2.$$

Підставивши у формулу дані табл. 68, дістанемо:

$$332 = a_0 + 3981a_1 + 55a_2;$$
$$147390 = 3981a_0 + 1762967a_1 + 20230a_2;$$
$$1723 = 45a_0 + 20230a_1 + 285a_2.$$

Розв'язавши цю систему рівнянь, матимемо такі значення параметрів:

a_0=3,536; a_1=0,072; a_2=0,301.

Отже, рівняння регресії, що характеризує вплив окремих факторів на урожайність зернових культур, має вигляд:

$$y_t = 3{,}636 + 0{,}072 + 0{,}301t.$$

Параметр a_1 цього рівняння показує, що з підвищенням енергозабезпеченості господарства зростає урожайність зернових культур на 0,072 ц. За рахунок інших факторів, які рівномірно змінюються протягом останніх 9 років, урожайність зернових культур щорічно зростала на 0,301 ц/га.

Якщо автокореляцію усунено, то різниці між фактичними та теоретичними рівнями ($Et = y - y_t$) мають бути незалежними. Цю гіпотезу перевіряють за допомогою коефіцієнта автокореляції, який розраховують з певним зсувом часу — лагом p. При $p = 1$ коефіцієнт автокореляції обчислюють за формулою

$$r_a = \frac{\sum E_t \cdot E_{t+1}}{\sum E_t^2}.$$

Коефіцієнт автокореляції коливається в межах від –1 до +1. Методику розрахунку коефіцієнта автокореляції наведено в табл. 69.

Підставивши підсумкові дані табл. 69 у формулу, дістанемо:

$$r_a = \frac{-13{,}26}{28{,}48} = -0{,}465.$$

При рівні значущості 0,05 і 7 ступенях свободи варіації ($n - 2 = 9 - 2$) критичне значення коефіцієнта кореляції дорівнює 0,666 (дод. 10). Оскільки фактичний коефіцієнт автокореляції менший за критичний, то це свідчить про низький ступінь кореляції залишкових величин та її випадковий характер.

Таблиця 69 *Розрахунок коефіцієнта автокореляції*

y	y_t	E_t	E_{t+1}	$E_t \cdot E_{t+1}$	E_t^2
32	33,5	+1,5	0,2	–0,30	2,25
35	34,8	0,2	–1,7	–0,34	0,04
34	35,7	+1,7	–0,4	0,68	2,89
36	36,4	–0,4	1,9	–0,76	0,16
39	37,1	1,9	–3,0	–5,70	3,61
35	38,0	–3,0	1,2	–3,60	9,0
40	38,8	1,2	0,3	0,36	1,44
39	38,7	0,3	3	0,90	0,09
42	39,0	3	–1,5	–4,50	9,00
Разом	332,0	0	0	–13,26	28,48

Контрольні питання

1. *Що таке кореляційний аналіз?*
2. *Що таке парна та множинна кореляції, прямий і обернений, прямолінійний і криволінійний кореляційні зв'язки?*
3. *Із яких послідовних стадій складається кореляційний аналіз?*
4. *Що таке кореляційне поле і яке його призначення?*
5. *Що таке рівняння регресії та лінії регресії?*
6. *Як визначаються рівняння регресії та коефіцієнт кореляції при парному прямолінійному зв'язку?*
7. *Як визначаються рівняння регресії й індекс кореляції при криволінійному зв'язку?*
8. *Як визначаються рівняння регресії та показники щільності зв'язку при множинній кореляції?*
9. *Як здійснюється статистична оцінка вибіркових показників зв'язку?*
10. *Що таке коефіцієнти еластичності і β -коефіцієнти?*
11. *Як визначається коефіцієнт кореляції рангів?*
12. *Як визначаються коефіцієнти взаємного узгодження й асоціації?*
13. *Які особливості кореляційного аналізу в рядах динаміки?*

Розділ 7
КОМПЛЕКСНЕ ЗАСТОСУВАННЯ МАТЕМАТИКО-СТАТИСТИЧНИХ МЕТОДІВ АНАЛІЗУ ДАНИХ

7.1 Загальне поняття про математико-статистичні методи

Поряд із традиційними статистичними методами аналізу даних при дослідженні реальних соціально-економічних явищ і процесів широко застосовуються математико-статистичні методи, виходячи з вітчизняної та зарубіжної методології. Комплексність застосування математико-статистичних методів передбачає найбільш повне розкриття суті, закономірностей і тенденцій розвитку конкретних явищ і процесів з метою більш адекватного відображення їх властивостей і особливостей, резервів і перспектив розвитку та шляхів удосконалення.

Ускладнення структури соціально-економічних явищ передбачає використання ряду методів класифікації та виділення однорідних груп, в основі побудови яких лежать міри близькості або метрики. Суть полягає в тому, що розподіл досліджуваних об'єктів або явищ у сукупності повинен підпорядковуватися нормальному закону розподілу, з тим, щоб отримати моделі, які дійсно будуть відображати якісно однорідні групи.

Нормальний розподіл важливий з багатьох причин. Розподіл багатьох статистик є нормальним або може бути отриманий з нормальних за допомогою деяких перетворень. Нормальний розподіл представляє собою одну з емпірично перевірених істин відносно загальної природи дійсності і його положення можуть розглядатися як один з фундаментальних законів природи. Точна форма нормального розподілу (характерна «колокообразна крива») визначається тільки двома параметрами: середнім і стандартним відхиленням.

Характерна властивість нормального розподілу полягає в тому, що 68 % усіх його спостережень лежать у діапазоні плюс-мінус одне стандартне відхилення від середнього, а діапазон ± 2 стандартних відхилення має 95 % значень. Іншими словами, при нормальному розподілі стандартизовані спостереження, менші –2 або більші +2, мають відносну частоту менше 5 %. Стандартизоване спостереження

означає, що від вихідного значення віднято середнє значення і результат поділено на стандартне середнє квадратичне відхилення (корінь з дисперсії).

При використанні пакету STATISTICA, можна визначити точні значення вірогідностей, пов'язаних з різними значеннями нормального розподілу, використовуючи Вірогідносний калькулятор. Наприклад, якщо задати значення випадкової величини, що має стандартне нормальне відхилення, $z = 4$, відповідний рівень вірогідності, визначений за пакетом STATISTICA буде менше 0,0001, оскільки при нормальному розподілі майже всі спостереження (тобто більше 99,99 %) потраплять у діапазон ± 4 стандартних відхилення.

Найбільшого поширення в практиці аналізу економічних явищ і процесів отримали:

- Кластерний аналіз.
- Метод головних компонент.
- Факторний аналіз.

7.2 Кластерний аналіз

Термін «кластерний аналіз» вперше введено в науковий оборот Трайоном (Tryon) у 1939 р. Оскільки це поняття включає в себе набір різних **алгоритмів** класифікації, дамо деяке пояснення терміна «алгоритм». У протилежність евристикам, які містять загальні рекомендації, засновані на статистичній очевидності і теоретичних міркуваннях, алгоритми є повністю визначеними кінцевими наборами кроків, операцій або процедур, які приводять до конкретного результату. Наприклад, за невеликим винятком, усі комп'ютерні програми, математичні формули є алгоритмами. Загальне питання, яке задають дослідники в багатьох областях, полягає в тому, як організувати спостережувані дані в наочні структури, тобто розгорнути таксономії.

Нехай існує n об'єктів, кожний з яких характеризується набором k - ознак. Необхідно розбити цю сукупність на однорідні групи. Отримані в результаті поділу групи називаються **кластерами**, а метод їх знаходження — **кластерним аналізом**. **Розмір кластерів** — це число об'єктів, що містяться в них.

Фактично кластерний аналіз є не стільки звичайним статистичним методом, скільки «набором» різноманітних алгоритмів розподілу об'єктів по кластерах. Існує точка зору, що на відміну від багатьох інших статистичних процедур, методи кластерного аналізу використовуються в більшості випадків тоді, коли дослідник не має будь-яких апріорних гіпотез відносно класів, але все ще перебуває в описовій стадії дослідження. Кластерний аналіз визначає найбільш можливе значиме рішення. Тому перевірка статистичної значимості тут не проводиться навіть у випадках, коли відомі p-рівні, як, наприклад, в методі K середніх.

Техніка кластерізації застосовується в найрізноманітніших областях. Хартиган (Hartigan, 1975 р.) зробив огляд багатьох опублікованих досліджень, що містять результати, отримані методами кластерного аналізу, в тому числі в області маркетингових досліджень,

археології, медицини, психіатрії. Кластерний аналіз широко застосовується і є корисним у тих випадках, коли необхідно класифікувати велику масу інформації на придатні для подальшої обробки групи.

Загальними методами кластерного аналізу є:

I. Об'єднання (деревоподібна кластеризація).

II. Двовходове об'єднання.

III. Метод К середніх.

Призначення алгоритму об'єднання або деревоподібної кластеризації полягає в об'єднанні об'єктів дослідження в досить великі кластери з урахуванням деякої міри схожості або відстані між об'єктами. Типовим результатом такої кластеризації є **ієрархічне дерево**.

Розглянемо **горизонтальну деревоподібну діаграму**. Діаграма починається з кожного об'єкта в класі (в лівій частині діаграми). Тепер поступовими, уявними, дуже малими кроками ви змінюєте ваш критерій визначення унікальних і неунікальних об'єктів. Іншими словами, ви понижуєте поріг, який веде до рішення про об'єднання двох або більше об'єктів в один кластер (дод. 1).

У результаті ви пов'язуєте разом все більшу і більшу кількість об'єктів і **об'єднуєте** (агрегатуєте) все більше і більше кластерів, які складаються з елементів, що все більше відрізняються між собою. У результаті, на останньому кроці всі об'єкти поєднуються разом. На деревоподібних горизонтальних діаграмах горизонтальні осі представляють відстань об'єднання (у **вертикальних деревоподібних діаграмах** відстань об'єднання представляють вертикальні осі). Так, для кожного вузла в графі (там, де формується новий кластер) можна побачити величину відстані, для якої відповідні елементи пов'язуються в новий єдиний кластер. Коли дані мають ясну структуру в термінах кластерів об'єктів, схожих між собою, тоді ця структура має бути відображена в ієрархічному дереві різними вітками. У результаті успішного аналізу методом об'єднання з'являється можливість виявити кластери (вітки) і інтерпретувати їх.

Найбільш важливим у кластерному аналізі вважається визначення однорідності об'єктів, які задаються введенням відстані між об'єктами

x_i і x_j (відстань $\rho(x_i, x_j)$). Вибір відстані є основним моментом дослідження, від якого залежать кінцеві варіанти кластеризації. Об'єкти будуть однорідними у випадку

$$\rho(x_i, x_j) \le \rho_{пор},$$

де $\rho_{пор}$ — задане порогове значення.

Найбільш поширеними вважаються такі міри відстані:

- Евклідова відстань;
- квадрат евклідової відстані;
- Хемінгова відстань;
- Манхеттенська відстань;
- відстань Чебишева;
- степенева відстань.

Вибір міри відстані визначається структурою ознакового простору та метою класифікації.

При використанні процедур кластерного аналізу розподіл об'єктів сукупності на якісно однорідні групи проводиться одночасно за великою кількістю ознак, але при дотриманні умови, що жодна ознака не виділяється за своєю значимістю так, що групування на її основі є головним. Особливістю кластерного аналізу є те, що відмінності між одиницями, що входять у виділену групу (кластер), незначні, а відмінності між групами — суттєві.

Об'єднання або метод деревоподібної кластеризації використовується при формуванні кластерів несхожості або відстані між об'єктами. Ці відстані можуть визначатися в одномірному або багатомірному просторі. Наприклад, якщо необхідно кластеризувати типи сільськогосподарських підприємств, то до уваги можна взяти розмір земельної площі, фондозабезпеченість, рівень рентабельності, собівартість виробленої продукції, ціну реалізації тощо. Найдоцільнішим шляхом визначення відстані між об'єктами в багатомірному просторі є метод **Евклідових відстаней**. Якщо ми маємо дво-

або тримірний простір, то ця міра є реальною геометричною відстанню між об'єктами в просторі (наче ця відстань між об'єктами виміряна рулеткою). Проте за алгоритму об'єднання не є важливим, чи є дійсними або деякими іншими похідними мірами відстані взяті для цього відстані що більш значимо для дослідника. Завданням дослідника є підбір правильного методу специфічних застосувань.

Евклідова відстань — це найбільш загальний тип відстані. Вона є геометричною відстанню в багатомірному просторі і визначається за формулою:

$$відстань\ (x, y) = \rho(x, y) = \left\{\sum_i (x_i - y_i)^2\right\}^{\frac{1}{2}}$$

або

$$\rho(x_i, x_j) = \sqrt{\left(x_i^{1,2,\ldots k} - x_j^{1,2,\ldots k}\right)^2}.$$

Квадрат евклідової відстані. Іноді може виникнути необхідність піднести до квадрата стандартну евклідову відстань, щоб надати більші ваги віддаленим один від одного об'єктам. Ця відстань визначається за формулою:

$$відстань\ (x, y) = \sum_i (x_i - y_i)^2.$$

Евклідова відстань (і її квадрат) визначається за вихідними, а не за стандартизованими даними. Це звичайний спосіб її визначення, який має певні переваги. Наприклад, відстань між двома об'єктами не змінюється при введенні в аналіз нового об'єкта, який може виявитися викидом. **Викиди** — це нетипові спостереження, які дуже вирізняються з-посеред інших. Проте на відстань можуть значно впливати відмінності між осями, за координатами яких вона визначається. Якщо,

наприклад, одна з осей виміряна в сантиметрах, а дослідник переводить її в міліметри, то кінцева евклідова відстань (або квадрат евклідової відстані), яка визначається за координатами, дуже зміниться, і, як наслідок, результати кластерного аналізу можуть значно відрізнятися від інших.

Хемінгова відстань використовується як міра відстані об'єктів, значення яких задано атрибутивними ознаками. Вона визначається за формулою:

$$\rho(x_i, x_j) = \sum_{i,j}^{1,2,\ldots,k} \left| x_i^{1,2,\ldots,k} - x_j^{1,2,\ldots,k} \right|$$

Манхеттенська відстань (відстань міських кварталів) є середньою різницею між координатами. У більшості випадків ця міра відстані призводить до таких же результатів, як і для звичайної відстані Евкліда, проте для неї вплив окремих великих відстаней (викидів, винятків) зменшується, оскільки вони не підносяться до квадрата. Манхеттенська відстань визначається за формулою:

$$\textit{відстань}\ (x, y) = \sum_i \left| x_i - y_i \right|$$

Відстань Чебишева — це відстань, яка використовується при необхідності визначення двох об'єктів як «різних», якщо вони відрізняються за будь-якою однією координатою (будь-яким одним виміром). Відстань Чебишева визначається за формулою:

$$\textit{відстань}\ (x, y) = \max \left| x_i - y_i \right| .$$

Степенева відстань — застосовується у випадках, коли необхідно прогресивно збільшити або зменшити вагу, що належить до розмірності, для якої відповідні об'єкти дуже відрізняються. Степенева відстань визначається за формулою:

$$\textit{відстань}\ (x, y) = \left(\sum_i \left| x_i - y_i \right|^p \right)^{\frac{1}{r}},$$

де r і p — параметри, що визначаються дослідником. Параметр p відповідальний за поступове зважування різниць за окремими координатами, параметр r — за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметри r і p дорівнюють двом, то ця відстань збігається з відстанню Евкліда.

На першому кроці, коли кожний об'єкт являє собою окремий кластер, відстані між цими об'єктами визначаються обраною мірою. Проте, коли пов'язують разом декілька об'єктів, виникає питання, як треба визначити відстань між кластерами? Іншими словами, необхідне **правило об'єднання (або зв'язку) для двох кластерів**.

Математична статистика розглядає такі правила об'єднання:

- Одинокий зв'язок (метод «найближчого сусіда»).
- Повний зв'язок (метод «найбільш віддалених сусідів»).
- Незважене попарне арифметичне середнє.
- Зважене попарне арифметичне середнє.
- Незважений попарний центроїдне усереднення.
- Зважений попарний центроїдний метод (медіана).
- Метод Варда.

Одинокий зв'язок (метод «найближчого сусіда»). При цьому методі відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами в різних кластерах. У результаті кластери, отримані таким чином, об'єднуються тільки окремими елементами, що випадково опинилися ближче інших один від одного. Це правило повинно об'єднувати об'єкти разом для формування кластерів, з тенденцією побудови у підсумку і подання кластерів довгими «ланцюжками».

Як альтернативу можна використовувати сусідів у кластерах, які найбільш віддалені від інших пар об'єктів і один від одного. Цей метод називається **методом повного зв'язку**.

Повний зв'язок (метод «найбільш віддалених сусідів»). При цьому методі відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах (тобто «найбільш віддаленими сусідами»). Цей метод спрацьовує дуже добре, коли об'єкти мають неоднакове, несхоже походження. Якщо ж кластери мають деяким чином подовжену форму або їх природний тип є «ланцюговий», то цей метод непридатний.

Незважене попарне арифметичне середнє. У цьому методі відстань між двома різними кластерами визначається як середня відстань між усіма парами об'єктів у них. Метод є ефективним, коли об'єкти в дійсності формують різні групи, проте він спрацьовує однаково добре і у випадках подовжених («ланцюгового» типу) кластерів. Відзначимо, що у своїй книзі Сніт і Сокел (Sneath, Sokal, 1973 р.) уводять абревіатуру **UPGMA** для посилання на цей метод як на *метод незваженого попарного арифметичного середнього — unweighted pair-group method using arithmetic averages.*

Зважене попарне арифметичне середнє. Метод ідентичний попередньому методу незваженого попарного середнього за винятком того, що при обчисленнях розмір відповідних кластерів (тобто число об'єктів, що містяться в них), використовується як ваговий коефіцієнт. Тому цей метод повинен використовуватися, коли передбачаються нерівні розміри кластерів. У книзі Сніта і Сокела (Sneath, Sokal, 1973 р.) вводиться абревіатура WPGMA для посилання на цей метод, як на метод зваженого попарного арифметичного середнього — weighted pair-group method usmg arithmetic averages.

Незважений попарний цетроїдний метод. В цьому методі відстань між двома кластерами визначається як відстань між їх центрами ваги. Сніт і Сокел використовують абревіатуру **UPGMC** для посилання на цей метод, як на *метод незваженого попарного цетроїдного усереднення — unweighted pair-group method using the centroid average.*

Зважений попарний цетроїдний метод (медіана). Цей метод ідентичний попередньому, за винятком того, що при обчисленнях використовуються ваги для врахування різниці між розмірами кластерів

(тобто числом об'єктів у них). Тому, якщо існують або передбачаються значні відмінності в розмірах кластерів, цей метод має переваги перед попереднім. Для посилання на нього як на метод зваженого попарного цетроїдного усереднення використовується абревіатура **WPGMC** — *weighted pair-group method using the centroid average.*

Метод Варда. Цей метод відрізняється від усіх інших методів, оскільки він використовує методи дисперсійного аналізу для оцінки відстаней між кластерами. Метод мінімізує суму квадратів (SS — від англійської Sum of Squares — сума квадратів) для будь-яких двох (гіпотетичних) кластерів, які можуть бути сформовані на кожному кроці. У цілому, метод є дуже ефективним, але за його допомогою складаються кластери малого розміру.

Двовходове об'єднання — потужний засіб розвідувального аналізу даних. Раніше цей метод обговорювався в термінах об'єктів, які повинні бути кластеризовані (див. «Об'єднання (деревоподібна кластеризація)»). В усіх інших видах аналізу питання, що цікавить дослідника, виражається в термінах спостережень і або змінних. Кластеризація, як по спостереженням, так і за змінними, може привести до досить цікавих результатів.

Дослідник збирає дані про різні характеристики (змінні) економічного стану сільськогосподарських підприємств (спостережень), що мають зерновий виробничий напрямок. Дослідник може кластеризувати спостереження (сільськогосподарські підприємства) для визначення кластерів господарств аналогічної спеціалізації. У той же самий час, дослідник може кластеризувати змінні для визначення кластерів змінних, які пов'язані схожими економічними показниками.

Звідси постає доцільне питанняпро можливість проведення кластеризації в обох напрямках? Модуль «*Кластерний аналіз*» містить ефективну двовходову процедуру об'єднання, яка дозволяє кластеризувати і спостереження, і змінні. Проте двовходове об'єднання використовується відносно рідко в обставинах, коли очікується, що і спостереження, і змінні мають однаково суттєве значення при побудові бажаних кластерів (дод. 2).

Так, повертаючись до попереднього прикладу, можна припустити, що дослідникові потрібно виділити кластери сільськогосподарських підприємств, схожих у відношенні до відповідних кластерів характеристик економічного стану. Проблема з інтерпретацією отриманих результатів виникає внаслідок того, що схожості між різними кластерами можуть бути викликані деякими відмінностями сукупності змінних. Тому отримані кластери є за своєю природою неоднорідними. У порівнянні з іншими зазначеними раніше методами кластерного аналізу двовходове об'єднання є рідко використовуваним методом.

Метод *K* середніх суттєво відрізняється від методів об'єднання (деревоподібна кластеризація) і двовходового об'єднання. Припустимо, що ми вже маємо гіпотези відносно кількості кластерів (за спостереженнями або за змінними). Ми можемо поставити завдання утворити рівно три кластери таким чином, щоб вони різнилися настільки, наскільки це тільки можливо. Саме такий тип задач вирішує алгоритм методу *K* середніх. В загальному вигляді метод К середніх створює *K* різних кластерів, розташованих на великих відстанях один від одного.

Приклад. У наведеному раніше прикладі з економічним станом сільськогосподарських підприємств (див. «Двовходове об'єднання»), дослідник може мати гіпотезу зі свого практичного досвіду, що досліджувані господарства попадають у три різні категорії. Далі він може побажати взнати, чи може його інтуїція бути підтверджена кількісно, тобто, чи насправді кластерний аналіз *K* середніх може дати три кластери підприємств, як передбачалося. Якщо це так, то середні різних мір економічних показників для кожного кластера будуть давати кількісний спосіб подання гінотез дослідника (наприклад, сільськогосподарські підприємства в кластері один мають високий параметр один, менший параметр два тощо).

Із точки зору розрахунків метод К середніх можна розглядати як дисперсійний аналіз «навпаки». Програма починається з *K* випадково відібраних кластерів, а потім ураховує належність об'єктів до

них, щоб, по-перше, мінімізувати зміни *усередині* кластерів і, по-друге, максимізувати мінливість *між* кластерами. Даний спосіб аналогічний методу «дисперсійний аналіз ANOVA навпаки», згідно з яким за допомогою критерію значущості в дисперсійному аналізі порівнюється міжгрупова варіацію з внутрішньогруповою при перевірці гіпотези про те, що середні в групах відрізняються одна від одної. В кластеризації методом *K* середніх програма переміщує об'єкти (тобто спостереження) з одних груп (кластерів) в інші для того, щоб отримати найбільш значимий результат при проведенні дисперсійного аналізу (ANOVA).

Інтерпретація результатів кластеризації методом *K* середніх проводиться таким чином. Як правило, коли результати кластерного аналізу методом К середніх отримані, можна розрахувати середні значення для кожного кластера за кожним показником, щоб оцінити, наскільки кластери відрізняються один від одного. В ідеалі ми повинні отримати сильно варіюючі середні для більшості, якщо не для всіх вимірювань, використаних в аналізі. Значення *F*-критерію статистики, отримані для кожного показника, є іншим індикатором того, наскільки добре відповідна ознака дискримінує кластери.

7.3 Ілюстративний приклад кластерного аналізу

Провести кластеризацію (класифікацію) шести сільськогосподарських підприємств, кожне з яких характеризується двома ознаками (табл. 70).

Таблиця 70 Вихідні дані для кластерного аналізу

Показник	1	2	3	4	5	6
Вартість валової продукції сільського господарства, тис. грн., x_1	2	4	5	12	14	15
Середньорічна вартість основних виробничих фондів, тис. грн., x_2	8	10	7	6	6	4

Представимо ці дані графічно на рис. 6.

Рис. 6 *Залежність між обсягом валової продукції та середньорічною вартістю основних фондів*

$$\rho(x_i, x_j) = \sqrt{\sum_{I=1}^{k} (x_{iI} - x_{jI})^2},$$

де I — ознаки, k — кількість ознак.

$$\rho_{11} = 0$$
$$\rho_{12} = \sqrt{(2-4)^2 + (8-10)^2} = \sqrt{8} = 2{,}83.$$

Розрахунки інших відстаней $\rho(x_i, x_i)$ аналогічні:

$\rho_{13} = 3{,}16$; $\rho_{14} = 10{,}19$; $\rho_{15} = 12{,}17$; $\rho_{16} = 13{,}60$;
$\rho_{23} = 3{,}16$; $\rho_{24} = 8{,}94$; $\rho_{25} = 10{,}77$; $\rho_{26} = 12{,}53$;
$\rho_{34} = 7{,}07$; $\rho_{35} = 9{,}06$; $\rho_{36} = 10{,}44$; $\rho_{45} = 2{,}00$;
$\rho_{46} = 3{,}61$; $\rho_{56} = 2{,}24$;

I. Принцип «найближчого сусіда»

$$R_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 2{,}83 & 3{,}16 & 10{,}19 & 12{,}17 & 13{,}60 \\ & 0 & 3{,}16 & 8{,}94 & 10{,}77 & 12{,}53 \\ & & 0 & 7{,}07 & 9{,}06 & 10{,}44 \\ & & & 0 & 2 & 3{,}61 \\ & & & & 0 & 2{,}24 \\ & & & & & 0 \end{pmatrix} \end{array};$$

$$\rho_{\min} = \rho_{45} = 2(S_1; S_2; S_3; S_{4,5}; S_6);$$

$$R_2 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4,5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4,5 & 6 \end{array} \\ \begin{pmatrix} 0 & 2,83 & 3,16 & 10,19 & 13,60 \\ & 0 & 3,16 & 8,94 & 12,53 \\ & & 0 & 7,07 & 10,44 \\ & & & 0 & 2,24 \\ & & & & 0 \end{pmatrix} \end{array};$$

$$\rho_{\min} = \rho_{4,5,6} = 2(S_1; S_2; S_3; S_{4.5,6});$$

$$R_2 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4,5,6 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4,5,6 \end{array} \\ \begin{pmatrix} 0 & 2,83 & 3,16 & 10,19 \\ & 0 & 3,16 & 8,94 \\ & & 0 & 7,07 \\ & & & 0 \end{pmatrix} \end{array}$$

$$\rho_{\min} = \rho_{12} = 2,83(S_{1,2}; S_3; S_{4,5,6});$$

$$R_2 = \begin{array}{c} \\ 1,2 \\ 3 \\ 4,5,6 \end{array} \begin{array}{c} \begin{array}{ccc} 1,2 & 3 & 4,5,6 \end{array} \\ \begin{pmatrix} 0 & 3,16 & 8,94 \\ & 0 & 7,07 \\ & & 0 \end{pmatrix} \end{array}$$

$$\rho_{\min} = \rho_{11,2,3} = 3.16(S_{1,2,3}; S_{4,5,6})$$

Таким чином, при проведенні кластерного аналізу за принципом «найближчого сусіда» ми отримали два кластери.

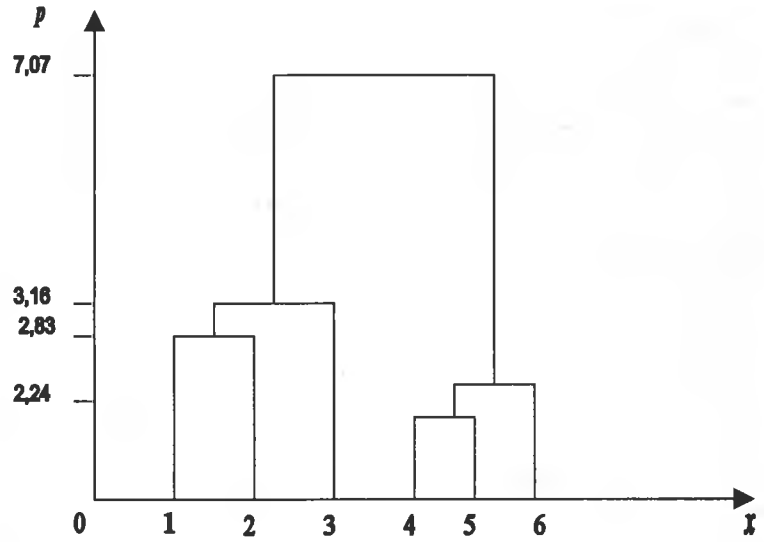


Рис. 7 *Дендрограма*

Моделі на основі результатів кластерного аналізу дозволяють досліджувати однорідні за основними економіко-технічними характеристиками і параметрами діяльності соціально-економічні об'єкти і процеси, а також стунінь їх ділової активності.

7.4 Метод головних компонент і факторного аналізу

У процесі поглиблення аналізу економічних структур — об'єктів і процесів — до розгляду включається все більше число ознак. При цьому необхідна оглядність («обозримость»). Закономірність розповсюджується на велику кількість зв'язків. Тому доцільно здійснювати класифікацію за декількома узагальнюючими ознаками, отриманими за допомогою **методу головних компонент** або **факторного аналізу**.

Передумовами використання цих методів є:

- Наявність сильно корелюючих ознак, наслідком чого є дублювання інформації.
- Слабка інформативність ряду факторних ознак.
- Можливість і доцільність агрегування декількох факторних ознак.

Скорочення розмірності досліджуваної сукупності методом головних компонент і факторного аналізу передбачає перехід до опису результативної ознаки меншою кількістю найбільш інформативних (з точки зору їх впливу на результативний показник), факторних ознак.

Метод головних компонент, що розглядається, як засіб зниження розмірності, також використовується для проведення класифікацій. Математичною моделлю, на якій заснований метод головних компонент,. такий багатомірний розподіл:

$$y = V_{j1}F_1 + V_{j2}F_2 + ... + V_{jp}F_p = \sum_{p=1}^{n} V_{jp}F_p\,, \quad j = 1,2,...,p\,,$$

де $V_{j1,2,...,p}$ — ваговий коефіцієнт загальної результативної ознаки (головної компоненти) на j-й змінній; $F_{1,2,...,p}$ — загальний фактор (головна компонента).

Суть методу полягає у виділенні лінійних комбіпацій вихідних факторних ознак, які мають максимально можливу дисперсію. При цьому, перша головна компонента володіє максимальною дисперсією і є нормованою лінійною комбінацією всіх можливих вихідних ознак, а друга — враховує максимальне значення дисперсії, що залишилась, і кореляційно не пов'язана з першою компонентою.

У цілому, якщо переходити до оцінки ступеня та напрямів зв'язків між досліджуваними ознаками на основі одночасного використання методів регресійного аналізу і головних компонент, рівняння головних компонент включає меншу їх кількість, ніж кількість факторних ознак, оскільки з подальшого дослідження повинні бути виключені головні компоненти, внесок яких у загальну дисперсію несуттєвий і становить менше 10 %.

Комплексне використання кореляційного аналізу та методу головних компонент виражається в розрахунку парних коефіцієнтів кореляції між вихідними (включеними в дослідження) ознаками та відповідними головними компонентами. На основі загальної оцінки значення парного коефіцієнта кореляції (за абсолютною величиною) і питомої ваги внеску (в процентному відношенні) кожної компоненти в загальну варіацію ознаки здійснюється відбір найбільш статистично суттєвих головних компонент.

До переваг методу головних компонент можна віднести:

1) у матриці значень компоненти розміщуються в спадаючому порядку власних значень, що сприяє класифікації ознак;
2) число компонент відповідає кількості вихідних факторних ознак;
3) головні компоненти не корелюють між собою, що є суттєвим при побудові регресійних моделей;
4) головні компоненти повністю зумовлюють варіацію вихідних факторних ознак.

Факторний аналіз полягає в переході від вихідної інформації до узагальнених факторів, які є результатом їх первісної агрегації і лінійної комбінації.

Головна модель факторного аналізу лінійна і має вигляд:

$$y_j = a_{1j}F_1 + a_{2j}F_2 + ... + a_{jp}F_p + d_jV_j,$$

де $F_1, F_2, ..., F_n$ — узагальнені фактори, що зумовлюють систематичну варіацію і кореляційний зв'язок між ними; a_{ij} — факторні навантаження (коефіцієнти регресії); V_i — характерні фактори, що враховують варіацію, яка не пояснюється загальними факторами.

Факторні навантаження оцінюють ступінь тісноти зв'язку між вихідними ознаками $x_1, x_2, ..., x_k$ та узагальненими факторами F_j. Зв'язок вважають суттєвим, якщо парний коефіцієнт кореляції більше або дорівнює за абсолютним значенням без урахування знаку 0,5.

У практичній діяльності внесок загального фактора в загальну дисперсію становить не менше 80 — 90 %.

Сукупне використання методів факторного та регресійного аналізів неможливе без урахування специфіки та відмінностей між ними. При використанні регресійного аналізу в модель не можуть бути включені всі змінні, які впливають на результативну ознаку, що веде до деякої втрати інформації. На основі реалізації факторного аналізу в модель включаються загальні фактори, які є реальним відображенням ряду економічно пов'язаних між собою вихідних змінних, які відтворюють і пояснюють їх властивості.

Перехід від загального числа факторних ознак до узагальнених факторів або головних компонент не є суттєвою втратою інформації. В окремих випадках загальні фактори можуть відображати властивості вихідних факторних ознак, які безпосередньо статистично не вимірювані і не здійснюють вплив на результат.

Моделі регресії на узагальнені фактори і головні компоненти не містять колінеарно пов'язаних ознак.

Проте, на практиці доцільно будувати моделі регресії на вихідні змінні, оскільки використання загальних факторів і головних компонент дуже ускладнює економічну інтерпретацію параметрів моделі.

Контрольні питання

1. *Які математичні методи застосовуються в практиці аналізу економічних явищ?*
2. *Що представляє собою нормальний розподіл і чим пояснюється його значення в аналізі однорідності сукупності?*
3. *Що включає в себе поняття «алгоритм»?*
4. *У чому полягає суть кластерного аналізу?*
5. *Що таке кластер і розмір кластеру?*
6. *Чи здійснюється перевірка статистичної значимості при кластерному аналізі?*
7. *Які загальні методи кластерного аналізу?*
8. *У чому полягає суть і завдання методу головних компонент?*
9. *У чому полягає суть і завдання факторного аналізу?*

Розділ 8
ВИКОРИСТАННЯ СИСТЕМИ «STATISTICA» В МАТЕМАТИЧНІЙ СТАТИСТИЦІ

8.1 Загальні відомості про систему «STATISTICA»

Мета даного розділу — ознайомити користувачів мультимедійного проекту StatSoft «STATISTICA»: искусство анализа данных» із сучасними технологіями аналізу даних на реальних прикладах. Він допоможе починаючим користувачам отримати базові знання з математико-статистичного аналізу економічних процесів з використанням програмних продуктів серії «STATISTICA».

Розробником системи «STATISTICA» є фірма *Stat.Soft., Inc.*, (США). «STATISTICA» — це інтегрована система аналізу й управління даними, інструмент розробки пропозицій користувачів у бізнесі, економіці, фінансах та інших областях. Всі аналітичні інструменти, що є в системі, доступні користувачу і можуть бути обрані за допомогою альтернативного інтерфейсу користувачів. Користувач може всебічно автоматизувати свою роботу, починаючи із застосування простих макросів для автоматизації трудомістких розрахунків до найпоглибленіших проектів, які включають також інтеграцію системи з іншими додатками або Інтернет.

Процедури системи «STATISTICA» мають високу швидкість і точність обчислень. Характерними рисами системи є:

- повний набір класичних методів аналізу даних (від основних методів статистики до сучасних);
- відповідає всім стандартам «Windows», що дозволяє зробити аналіз високоінтерактивним;
- дані системи «STATISTICA» легко конвертувати в різні бази даних і електронні таблиці;
- підтримує високоякісну графіку, яка дозволяє ефективно візуалізувати дані та проводити графічний аналіз;
- є відкритою системою: містить мови програмування, які дозволяють розширювати систему, запускати її з інших Windows-додатків, наприклад, з Excel.

«STATISTICA» складається з набору модулів, у кожному з яких зібрані тематично пов'язані групи процедур. При переключенні модулів можна залишати відкритим тільки одне вікно додатка «STATISTICA», або ж всі викликані раніше модулі, оскільки кожний з них може виконуватися в окремому вікні (як самостійний додаток Windows).

Настройка системи «STATISTICA». У системі передбачена можливість настройки багатьох характеристик і інтерфейса програми у відповідності до потреб користувачів. Можна змінити, наприклад, процес запуску, а саме, змінити встановлений за умовчанням повноекранний режим, змінити вид стартової панелі, панелі інструментів, таблиць з даними та інші параметри.

Настройка загальних параметрів системи. Настройку загальних параметрів системи можна змінити в будь-який момент роботи з програмою. Ці параметри визначають:

- загальні аспекти поведінки програми (максимізація вікна «STATISTICA» при запуску, Робочі книги, автоматичні зв'язки між графіками та даними, багатозадачний режим тощо);
- режим уводу (наприклад, автоматичне роздрукування таблиць і графіків, формати звітів, буферизація);
- загальний вид вікна додатка (значки, панелі інструментів тощо);
- вид вікон документів (колір, шрифти).

Кожний із цих параметрів можна настроїти у відповідному вікні, доступ до якого здійснюється через меню *Сервис*. Усі загальні параметри можуть бути настроєні незалежно від типу вікна документа (наприклад, таблиця або графік), яке активне на даний момент часу.

8.2 Введения даних в систему «STATISTICA»

Вихідні дані в системі «STATISTICA» організовані у вигляді таблиць. Таблична структура даних «STATISTICA» дозволяє природно, натурально відобразити більшість реальних даних.

Електронна таблиця складається з:

1) строк (рядків) — *Cases — Спостереження*;
2) стовпчиків — *Variables — Змінні*.

Наприклад, у сільському господарстві спостереження — це сільськогосподарські підприємства, змінні — загальна земельна площа, площа й урожайність окремих сільськогосподарських культур, собівартість, ціна реалізації, рівень фондозабезпеченості, рентабельності тощо.

Порядок створення таблиць, уведення даних у модулях системи «STATISTICA» розглянемо на прикладі модуля *Основні статистики і таблиці.*

Для того, щоб створити таблицю з даними необхідно зробити такі дії:

1. Запустіть програму «STATISTICA».
2. Відкрийте меню *Статистичних модулів* («STATISTICA» Module Switcher).
3. Оберіть з меню модуль *Основні статистики і таблиці* та клацніть по ньому мишею.
4. Тепер ви перебуваєте в модулі *Основні статистики і таблиці*, в якому можете обрати будь-яку статистичну процедуру, яка входить до цього модуля. Але оскільки в нас ще не створена база даних, клацніть мишею по кнопці *Вихід (Cancel)*.

Отже, ми перебуваємо в робочому вікні модуля *Основні статистики і таблиці* системи «STATISTICA»:

1. В основному робочому вікні системи підведіть курсор миші до рядка меню *Файл* і натисніть лівою кнопкою.

2. У випадаючому меню оберіть команду *Створити дані*. На екрані комп'ютера одразу ж з'явиться вікно *Створення даних*. У цьому вікні можна ввести ім'я файла. Файл може бути названий російською або українською мовами, але з цілого ряду причин доречно використовувати англійські імена.
3. Перемістіть курсор миші в поле *File name — Ім'я файла* та наберіть з клавіатури необхідне ім'я.
4. Натисніть клавішу *Enter* на клавіатурі або кнопку *Save* — програма утворить порожню таблицю, яка матиме 10 строк і 10 стовпчиків.
5. Можна збільшити або зменшити кількість рядків або стовпчиків цієї таблиці. Для цього використовуйте кнопки *Змінні/Спостереження* на панелі інструментів. Наприклад, натисніть кнопку *Спостереження*. Після натискання кнопки на екрані з'явиться меню, яке передбачає такий набір для спостережень таблиці: *Добавити, Перемістити, Копіювати, Видалити, Ввести імена спостережень*.
6. Оберіть, наприклад, пункт *Добавити*, двічі натиснувши ліву кнопку миші. Відкриється вікно, в якому можна задати кількість спостережень, добавляємих у таблицю (наприклад, 2).
7. Натисніть *ОК* — і кількість рядків (спостережень) у таблиці збільшиться на 2, тобто стане дорівнювати 12.
8. Аналогічно змініть число змінних у таблиці. Натисніть кнопку Змінні на панелі інструментів. За допомогою курсору миші у випадаючому меню оберіть пункт *Добавити*. На екрані з'явиться вікно, в якому задайте число змінних, які необхідно добавити (наприклад, 1). Натисніть *ОК* — і кількість стовпчиків (змінних) у таблиці збільшиться на 12, тобто стане дорівнювати 11.
9. Натисніть кнопку *Спостереження* й оберіть пункт меню *Імена*. На екрані з'явиться діалогове вікно, в якому можна визначити, скільки символів у таблиці буде зарезервовано для імен спостережень. Розширити поле для імен спостережень можна також за допомогою миші.

Таким чином, ми здійснили перший крок для подальших розрахунків — утворили електронну таблицю, яка має 11 стовпчиків і 12 рядків, а також місце для введення імен спостережень.

Тепер необхідно ввести назву таблиці (її заголовок) та імена змінних. Працюємо, використовуючи мишу. Головний принцип — двічі клацаючи мишею по полях заголовків, відкриваються діалогові вікна, які дозволяють вводити заголовки, описувати змінні тощо:

1. Двічі натисніть мишею на верхньому пустому рядку таблиці, який лежить над змінними. У вікні, яке з'явиться, введіть з клавіатури заголовок таблиці і натисніть *ОК*.
2. Аналогічно редагуються імена змінних і спостережень. Наприклад, щоб увести імена необхідно двічі клацнути мишею в полі. Ім'я спостереження і у вікні, яке з'явиться, ввести імена сільськогосподарських підприємств.
3. Для того, щоб описати змінну, необхідно двічі клацнути мишею на імені, наприклад, після натискання по заголовку *змінної 1 (Var 1)* відкриється вікно, в якому можна задати її ім'я або замінити ім'я на інше, формат змінної, зв'язок тощо.
4. Числові дані вносяться у таблицю безпосередньо з клавіатури, використанням стрілки пересування курсору. Поставте курсор у потрібне місце у таблиці та введіть числові дані.
5. Текстові значення вводяться інакше. Підведіть курсор до змінної з текстовим значенням і двічі натисніть мишею. У чарунці з'явиться код 999 — це код пропущених значень. Зітріть код за допомогою клавіши DEL на клавіатурі. Потім уведіть потрібне текстове значення..

Таким чином, працюючи в модулі *Основні статистики і таблиці*, ви побудували таблицю і ввели в неї дані. Подібним способом можна ввести дані в будь-якому модулі системи «STATISTICA». З точки зору загальних можливостей по управлінню даними, модулі системи однакові.

В системі «STATISTICA» мається спеціальний модуль *Управління даними (Data management)*, який містить розширені можливості,

що дозволяють швидко створити електронну таблицю, об'єднати дві таблиці, вирізати частину таблиці, відсортувати спостереження за будь-якою ознакою: наприклад, розташувати підприємства в алфавітному порядку або проран жирувати їх за розміром земельної площі тощо.

8.3 Найпростіші описові (дескриптивні) статистики

Змінні (англійський термін variable) — це те, що можна вимірювати, контролювати або чим можна манінулювати в дослідженнях. Іншими словами, змінна — це те, що варіює, змінюється, а не є постійним. Наприклад, вимірюючи урожайність сільськогосподарських культур, ми отримає різні значення в різних господарствах або різні значення в одному і тому самому господарстві, але в різні роки.

Змінні різняться характеристиками, а саме тим значенням, яку вони відіграють в дослідженнях., типом вимірювань тощо.

Оскільки значення змінних не постійні, потрібно навчитися описувати їх мінливість. Для цього є описові або дескриптивні статистики: мінімум, максимум, середнє, дисперсія, стандартне відхилення, медіана, квартилі, мода тощо. Ідея цих статистик проста: замість того, щоб розглядати *всі* значення змінної, а їх може бути досить багато, спочатку слід продивитися описові статистики. Вони дають загальне уявлення про значення, які приймає змінна.

Мінімум і максимум — це мінімальне і максимальне значення змінної.

Середнє — сума значень змінної, поділена на *n* (кількість значень змінної).

Дисперсія вибірки (від англійського variance) і стандартне відхилення (від англійського *standard deviation*) — міри *мінливості* змінної, які знайшли найбільш широке застосування. Дисперсія змінюється від 0 до нескінченності. Крайнє значення 0 означає відсутність коливання, коли значення змінної постійні.

Стандартне відхилення визначається як корінь квадратний з дисперсії. Чим вища дисперсія або стандартне (середнє квадратичне) відхилення, тим сильніше розсіяні значення змінної відносно середнього. Часто стандартне відхилення — більш зручна характеристика, оскільки показується в тих самих одиницях вимірювання, що й вихідна змінна.

Медіана вибірки (термін вперше введено Гальтоном, 1882 р.) розподіляє вибірку на дві рівні частини. Половина значень змінної лежить

нижче медіани, половина — вище. Медіана дає загальне уявлення про те, де зосереджені значення змінної, іншими словами, де знаходиться її центр. Медіана має таку властивість: сума *абсолютних* відстаней між точками вибірки і медіаною *мінімальна*. В деяких випадках, наприклад, при описі доходів населення, медіана більш зручна, ніж середнє значення ознаки.

Квартилі (термін вперше введено Гальтоном, 1882 р.) представляють собою значення, які ділять дві половини вибірки (розбиті медіаною) ще раз навпіл. Таким чином, медіана і квартилі ділять діапазон значень змінної на чотири рівні частини. Розрізняють верхню квартиль, яка більше медіани і ділить на дві рівні частини верхню частину вибірки (значення змінної більше медіани), і нижню квартиль, яка менше медіани і ділить навпіл нижню частину вибірки.

Верхню квартиль позначають символом 75 % — це означає, що 75 % значень змінної менше верхньої квартилі. Нижню квартиль часто позначають символом 25 %, це означає, що 25 % значень змінної менше нижньої квартилі.

Мода (термін вперше введено Пірсоном, 1894 р.) представляє собою значення змінної, яке повторюється максимально часто (має найбільшу частоту).

З описовими статистиками пов'язані статистичні графіки.

Які з описових статистик належить обирати? Середнє і медіана оцінюють положення центра вибірки, навколо якого групуються значення змінної. Маючи багато позитивних властивостей, середня чуттєва до викидів, які вносять в неї зрушення. Щоб усунути цей недолік, використовують зважену середню.

Оскільки медіана є середньою точкою варіаційного ряду, вона не така чуттєва до викидів. Якщо розподіл несиметричний (зміщений вліво або вправо), то медіана і міжквартильний розмах можуть дати більше інформації про те, в якій області сконцентровано спостереження.

Якщо медіана менше середнього значення, то розподіл зміщено вправо. Якщо медіана більше середнього, то розподіл зміщено вліво.

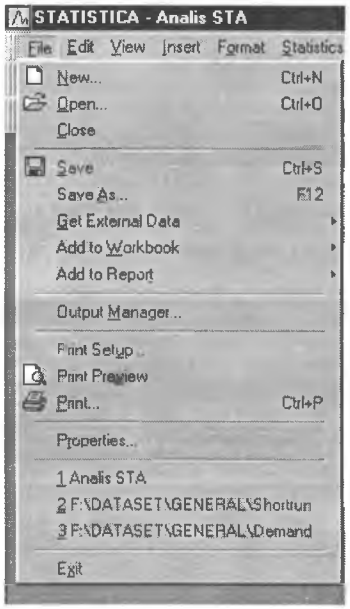
Як правило, при умові, що розподіл має одну моду, схема вибору статистик наступна. Якщо дані категоризовані, використовуйте моду.

Якщо не всі вихідні значення змінної представляють інтерес для дослідника, розподіл несиметричний і маються виброси, використовуйте медіану. В протилежному випадку використовуйте середнє значення ознаки.

Версія «STATISTICA» 6.0

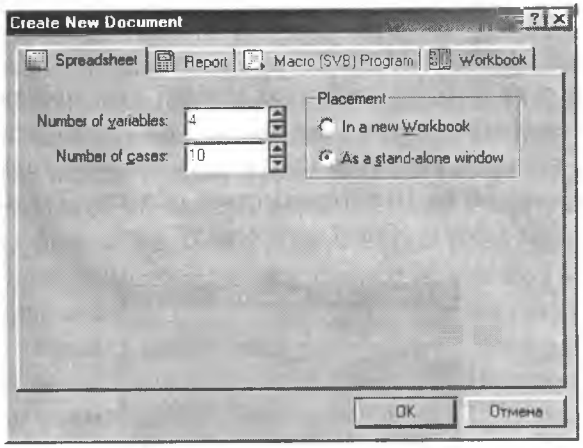
Введемо в електронну таблицю вихідні дані про якість ґрунтів (балів), внесення мінеральних добрив на 1 га озимої пшениці (т/га), урожайність озимої пшениці (ц/га), затрати людино-годин на виробництво 1 ц продукції по 10 підприємствах області. Для цього:

1. Із меню *File (Файл)* обираємо команду *New*.

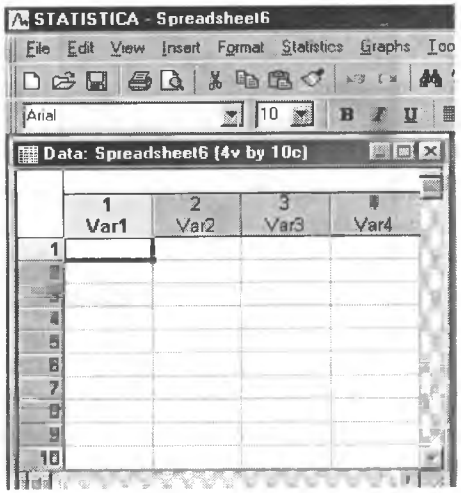


2. Зробимо у таблиці стільки змінних і випадків (спостережень), скільки необхідно. Для вихідних даних необхідно чотири змінних: якість ґрунтів, кількість добрив, урожайність, затрати праці на виробництво 1 ц продукції. В діалоговому вікні, що з'являється на екрані,

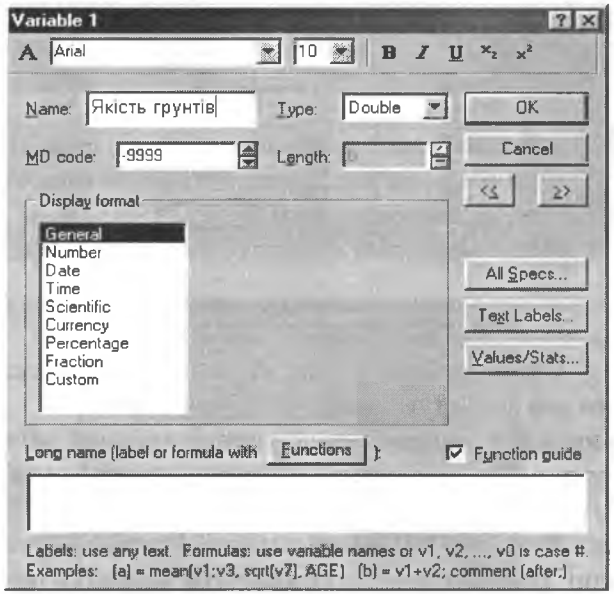
Create New Document в строчці *Number of variables* задайте число 4 (4 графи). В строчці *Number of cases* задайте число 10 (10 рядків).



3. Натисніть кнопку *OK*. «STATISTICA» утворить порожню електронну таблицю, яка появиться на екрані.

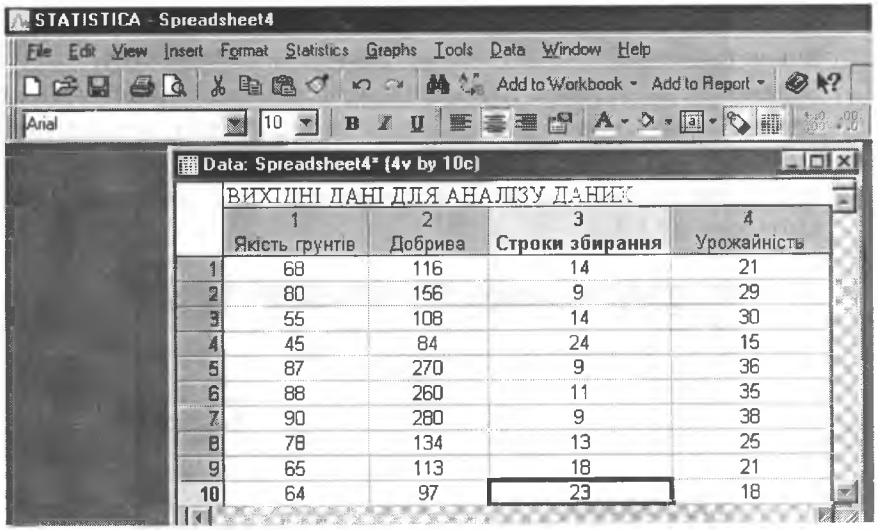


4. Необхідно підготувати таблицю до вводу даних, заголовок таблиці і імена змінних. Для цього двічі клацніть мишею на білому полі у таблиці під словами: Data, в якому можна задавати заголовок таблиці і додаткову інформацію про дані. Введемо заголовок таблиці: ВИХІДНІ ДАНІ ДЛЯ АНАЛІЗУ ДАНИХ.
5. Задамо імена змінних. Двічі натисніть на імені змінної VAR1 в електронній таблиці. На екрані з'явиться вікно специфікації змінної VAR1. В полі Name (Ім'я) напишіть з клавіатури: ГРУНТИ. Довжина імені не повинна перевищувати 8 символів. Натисніть кнопку *OK*.



Теж саме зробіть для змінної VAR2, яку назвіть добрива; змінній VAR3 дайте ім'я СТРОКИ ЗБИРАННЯ; змінну VAR4 назвіть УРОЖАЙНІСТЬ.

Тепер таблиця готова до того, щоб в неї ввести дані. Оскільки даних небагато, їх можна ввести з клавіатури.

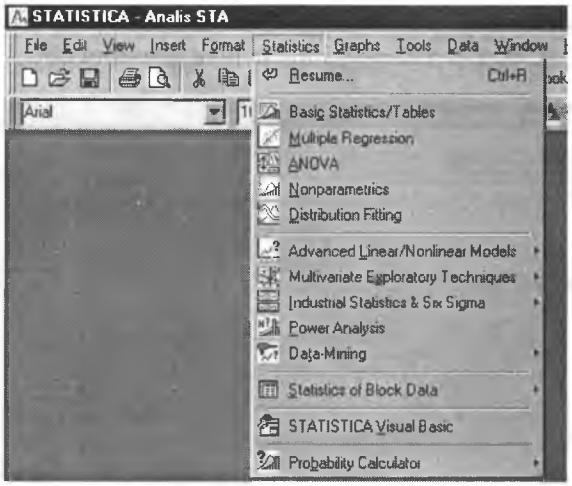


	1 Якість грунтів	2 Добрива	3 Строки збирання	4 Урожайність
1	68	116	14	21
2	80	156	9	29
3	55	108	14	30
4	45	84	24	15
5	87	270	9	36
6	88	260	11	35
7	90	280	9	38
8	78	134	13	25
9	65	113	18	21
10	64	97	23	18

6. Для збереження даних натисніть мишею на кнопку *Save As... (Збереження файла даних)* або наберіть з клавіатури *CTRL+S*. Створений файл буде збережено і він завжди буде доступним для користування.

На прикладі створеної електронної таблиці «Вихідні дані для аналізу даних» розглянемо **порядок визначення описових статистик**:

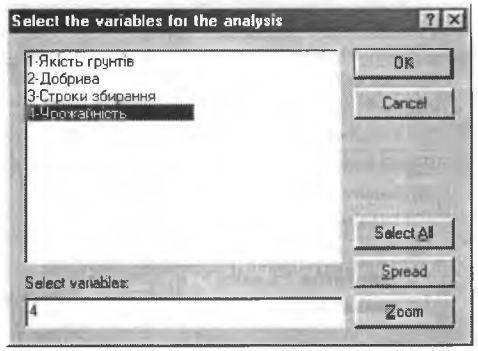
1. Виділіть будь-яке значення змінної *Урожайність* у таблиці «Вихідні дані для аналізу», для чого клацніть по ньому мишею.
2. У випадаючому меню *Statistics* натисніть кнопку *Basic Statistics (Швидкі основні статистики)* на панелі інструментів електронної таблиці.



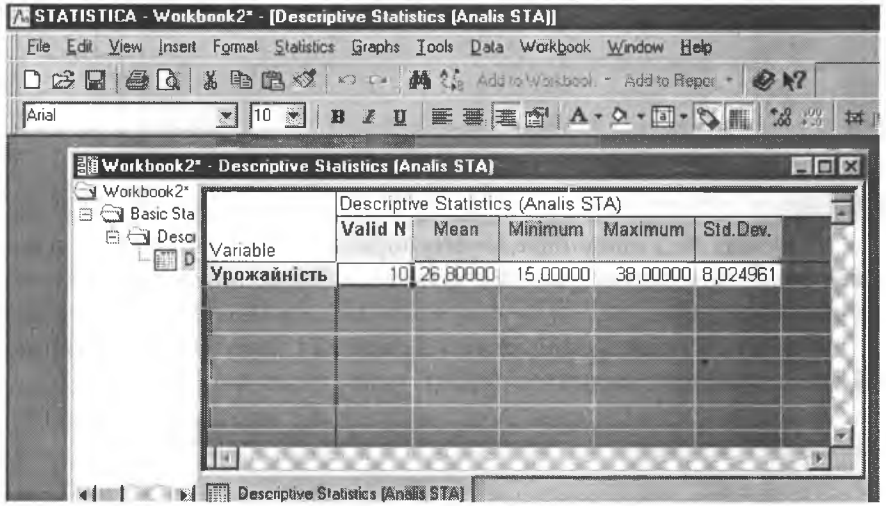
3. Кнопка *Швидкі основні статистики (Basic Statistics)* дозволяє швидко розрахувати основні статистики для однієї або цілого списку змінних. Всі види аналізу можуть бути проведені для згрупованих даних. Після натискання кнопки на екрані з'явиться випадаюче меню.

4. У випадаючому меню висвітіть верхню строчку *Descriptive statistics (Описові змінної)* і натисніть клавішу *OK*.

5. В режимі *Quick* вкажіть курсором на клавішу *Summery: Descriptive statistics*. На екрані з'явиться робоче вікно *Select the variables for the analysis*, в якому ви побачите перелік всіх досліджуваних змінних: якість ґрунтів, кількість внесених добрив строки збирання, урожайність.



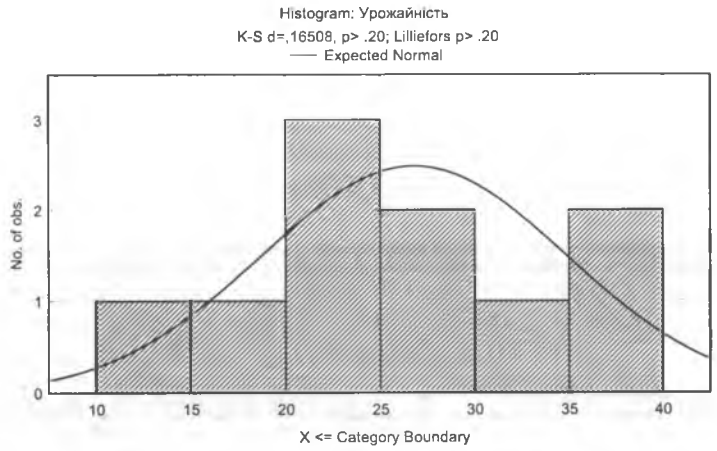
6. В пункті 1 на початку аналізу ми задали для дослідження показник урожайності. Тому у вікні, що перед вами на екрані також виділіть четвертий показник «*Урожайність*». Натисніть клавішу *OK*.
7. На екрані з'явиться електронна таблиця з основними описовими статистиками для обраної змінної *Урожайність*.



Отже, в результаті аналізу урожайності по 10 сільськогосподарських підприємствах ми отримали:

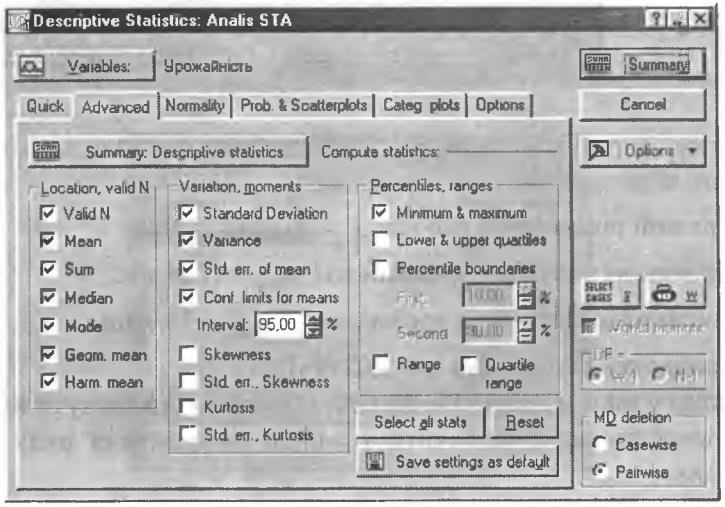
- середній рівень урожайності $\bar{x} = 26{,}8$ ц/га;
- мінімальне значення урожайності $x_{\min} = 15.0$ ц/га;
- максимальне значення урожайності $x_{\max} = 38{,}0$ ц/га;
- стандартне відхилення $\sigma = 8{,}024961$ ц/га.

Якщо у вікні *Descriptive statistics* (пункт 4) натиснути клавішу *Histograms*, зможемо побачити графічне зображення ряду «*Урожайність*».

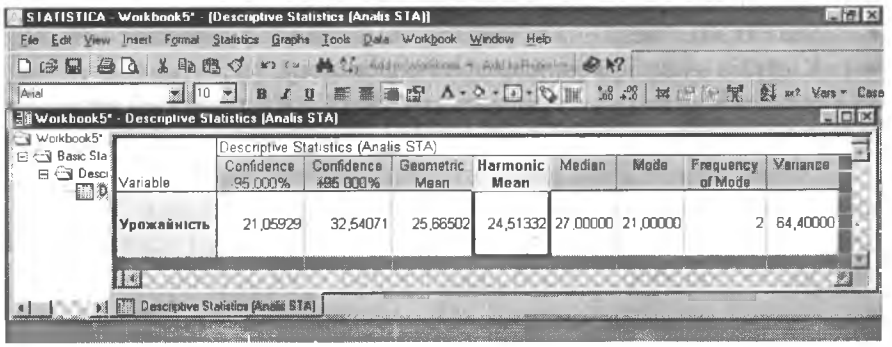


В пункті № 5 вище наведеного порядку дій ми працювали в режимі *Quick* (швидкому режимі), який дає обмежений перелік статистичних характеристик.

Перейти в розширений режим *Advanced* і зазначимо в ньому необхідні характеристики:



Натиснувши на клавішу Summary ми отримаємо дещо іншу, розширену електронна таблиця з основними описовими статистиками для обраної змінної *Урожайність*:



Таблиці результатів відрізняються від таблиці з вихідними даними. Цей спеціальний тип таблиць в «STATISTICA» носить назву **scrollsheets**.

Прокручуючи електронну таблицю результатів, зліва направо розташовані такі описувані статистики змінної ЗАТРАТИ:

- Valid N — справжнє (вихідне) число спостережень змінної ЗАТРАТИ (число випадків без пропусків);
- Mean — вибіркова Середня;
- Confid — 95.000 % — нижня межа 95 % довірчого інтервалу для середнього значення;
- Confid + 95.000 % — верхня межа 95 % довірчого інтервалу для середнього значення;
- Sum — сума (сума значень змінної ЗАТРАТИ);
- Minimum — мінімум (мінімальне значення змінної ЗАТРАТИ);
- Maximum — максимум (максимальне значення змінної ЗАТРАТИ)
- Range — розмах варіації (різниця між мінімальним і максимальним значенням змінної ЗАТРАТИ);

- Variance — вибіркова дисперсія
- Std.Dev. — стандартне відхилення;
- Std.Err. — стандартна помилка;
- Skewness — вибірковий коефіцієнт асиметрії;
- Std.Err.Skewness — стандартна помилка коефіцієнта асиметрії;
- Kurtosis — вибірковий коефіцієнт ексцесу;
- Std.Err.Kurtosis — стандартна помилка коефіцієнта ексцеса.

8.4 Статистичні графіки

Графічні файли системи «STATISTICA» містять не тільки зображення, а і всю інформацію для настройки графіка і аналізу даних. Тут записані всі подані на графіку дані, їх зв'язки. Записані в такому форматі графіки можна в подальшому відкрити в будь-якому з модулів системи «STATISTICA» для продовження настройки і аналізу даних. Крім того, їх можна роздрукувати в пакетному режимі за допомогою команди *Печать* файлів з випадаючого меню *Файл*.

Дані, представлені на графіках системи, можна безпосередньо продивлятися і змінювати незалежно від їх типу в *Редакторі даних графіка*. Це можуть бути вихідні дані, частини таблиць результатів або ряд розрахованих значень. Для кожного графіку утворюється пов'язане з ним «дочірнє» вікно *Редактора*, яке закривається разом зі своїм графічним вікном. *Редактор* організовано у вигляді груп стовпчиків, які представляють окремі залежності даного графіка.

В будь якому діалоговому вікні кожного типу графіків є кнопка Галерея графіків «STATISTICA», за допомогою якої відкривається діалогове вікно *Галерея графіків* «STATISTICA». Звідси швидко і легко викликаються всі графіки, для чого необхідно виділити назву необхідного типу графіка і двічі натиснути на ньому (або натиснути кнопку *ОК*).

Статистичні графіки пропонують сотні заздалегідь визначених типів графічних зображень, які включають аналітичне узагальнення статистичних даних. Вони викликаються з діалогового вікна *Галерея графіків*, яке відкривається за допомогою однойменної кнопки на панелі інструментів або з випадаючого меню *Графіка*.

При побудові таких графіків використовуються значення безпосередньо з файла даних, які не залежать від змісту поточної таблиці, виділення блоків і положення курсору. При цьому пропонуються або стандартні методи графічного аналізу даних (різні графіки розсіювання значень, гістограми, графіки середніх значень), або стандартні аналітичні методи досліджень (графіки нормальної щільності

розподілу, вірогідносні графіки з виключеним трендом, або графіки довірчих інтервалів ліній регресії). При побудові *статистичних графіків* програма враховує умови вибору і ваги спостережень.

Найбільш широко використовувані типи *статистичних графіків* (що викликаються з меню Графіка) представлені в меню *Швидкі статистичні графіки*. Ці списки графіків не мають такий широкий спектр можливостей, як меню *Статистичні графіки*, але на відміну від останніх спрощують і прискорюють процедуру побудови графіка. Швидкі статистичні графіки:

- викликаються із контекстних меню або з панелі інструментів будь-якої таблиці (як правило, не потребують звернення до випадаючих меню або діалогових вікон);
- не потребують від користувача вибору змінних (цей вибір визначається поточним положенням курсору у таблиці) і проміжної настройки параметрів (формат відповідних графіків визначається за умовчанням).

При виборі пункта *Швидкі статистичні графіки* (за допомогою кнопки на панелі інструментів або з випадаючого меню *Графіка*) з'являється меню вибору статистичного графіку для поточної змінної таблиці, тобто для тієї, на яку в даний момент вказує курсор. Якщо курсор не вказує ні на яку із змінних, то перед побудовою будь-якого графіка з меню *Швидкі статистичні графіки* буде запропоновано вибрати змінну із списка.

Будь-які зміни параметрів графіка в «STATISTICA» здійснюється з активного графічного вікна після відображення графіка на екрані. Як правило, спочатку має сенс побудувати графік, прийняв значення параметрів за умовчанням, а потім вже вносити різні зміни.

Проте в деяких випадках, коли побудова графіка займає досить багато часу (при обробці великих наборів даних), можна утручатися в цей процес, щоб зробити необхідні настройки. Припинити побудову графіка можна одним натисканням клавіши або миші в любому місті екрана, а потім продовжити його після вводу необхідних змін.

Найбільш простий і швидкий спосіб зміни параметрів будь-якого елемента графіка — подвійне натискання на ньому кнопкою миші. Крім того, за допомогою одного натискання правою кнопкою миші на даному об'єкті можна викликати відповідне йому контекстне меню.

Наприклад, при натисканні правою кнопкою миші на одній із осей графіка з'явиться контекстне меню, в якому пропонується вибір варіантів настройки для даної осі.

Всі графіки зберігають зв'язки з таблицею вихідних даних, за якими вони побудовані. При цьому, якщо оновлення не відбувається вручну і зв'язки не відмінені, графік автоматично оновлюється при зміні вихідних даних. Для управління зв'язками має ся спеціальне діалогове вікно *Зв'язку даних і графіка*. Воно викликається з випадаючого меню *Графіка*.

Тут можна встановити автоматичний режим зв'язку, коли графік автоматично оновлюється при зміні даних, за якими він побудований. Можна також задати режим *Вручну* або тимчасово заблокувати зв'язок. Крім того, можна встановити режим *Зв'язок з поточним файлом* даних і побудувати такий же графік або серію графіків для інших файлів даних.

Графіки і рисунки можуть бути збережені в графічному форматі «STATISTICA». Для цього використовують команди *Зберегти* або *Зберегти як...* з випадаючого меню Файл. Саме цей формат рекомендується для запису графічного файла, якщо передбачається в подальшому знову відкривати його в системі «STATISTICA». На відміну від інших графічних форматів формат «STATISTICA» зберігає не тільки саму картинку, але й *Редактор даних графіка* з усіма поданими на графіку даними, всі аналітичні параметри, а також інші параметри, що дозволяють в подальшому продовжити аналіз графічних даних. Цей формат найбільш зручний при пов'язуванні або перетворенні графіка в інший графік «STATISTICA». Збережені в даному графічному форматі файли

можна роздрукувати в пакетному режимі за допомогою команди *Друк файлів* з випадаючого меню *Файл*.

Окремі параметри, які визначають джерело даних, необхідно задати до побудови графіка, наприклад, змінні, значення міток, імена спостережень, мітки осей.

8.5 Простий кореляційний аналіз

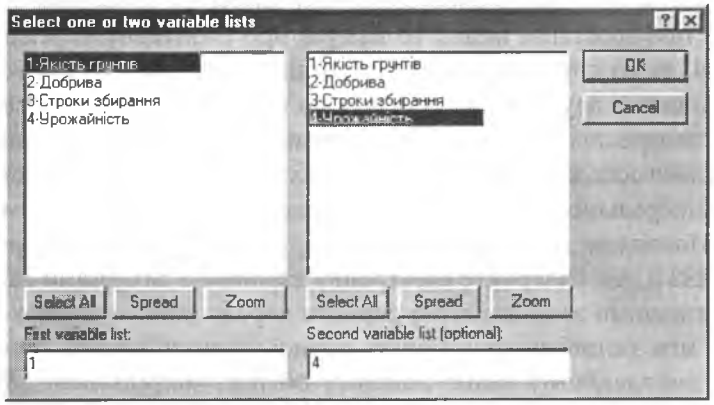
Кореляція, або, точніше, коефіцієнт кореляції, являється мірою залежності двох величин. Значення коефіцієнта кореляції знаходиться в межах від –1 і +1. Якщо спостерігається тенденція зростання однієї величини при зростанні іншої, то кажуть про позитивну кореляційність величин, якщо спостерігається тенденція збільшення однієї величини при зменшенні другої, то кажуть про від'ємну кореляційність величин.

В попередніх розділах ми створили файл з вихідними даними по 10 сільськогосподарських підприємствах (якість ґруптів, кількість внесених мінеральних добрив, строки збирання, урожайність озимої пшениці). Покажемо на основі цих даних, як обчислюються кореляції в «STATISTICA». Відкрийте електронну таблицю з вихідними даними.

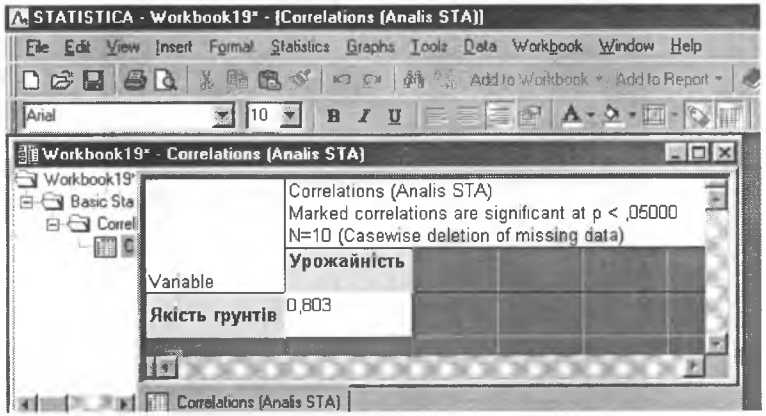
Визначимо за допомогою парного кореляційного аналізу тісноту зв'язку між змінними урожайність озимої пшениці і якість ґрунтів.

1. У випадаючому меню *Statistics* оберіть модуль *Basic Statistics (Швидкі основні статистики)*. У вікні, що з'явилося, виберіть *Correlation matrices*. Натисніть клавішу *OK*.

2. У робочому вікні *Product-Moment and Partial Correlations* в режимі *Quick* натисніть кнопку *Summary Correlation matrix*. На екрані з'явиться вікно Select one or to variable lists, в якому необхідно вибрати залежну і незалежну змінні. Натисніть клавішу *OK*.

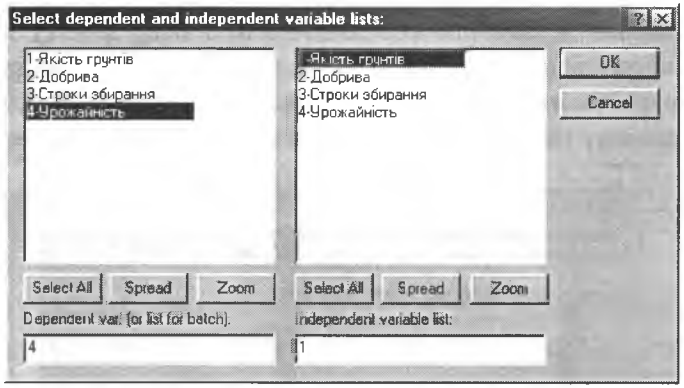


3. На екрані з'явиться електронна таблиця, в якій наведено коефіцієнт парної кореляції між урожайністю і якістю ґрунтів ($r_{y_x} = 0{,}803$).

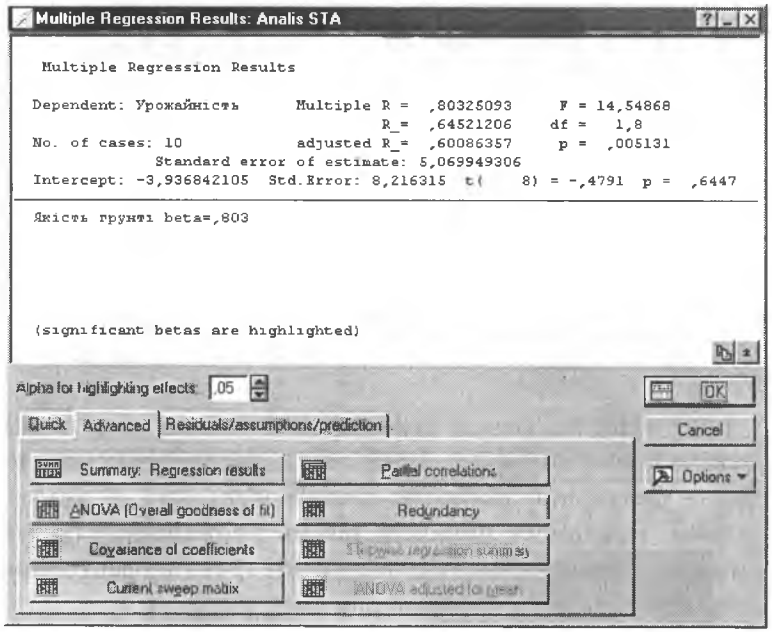


4. Для більш детального кореляційного аналізу зв'язку між урожайністю озимої пшениці і якістю ґрунтів (для розрахунку параметрів рівняння регресії, детермінації тощо) необхідно у випадаючому меню *Statistics* обрати модуль *Multiple Regression*. В режимі *Quick* натиснути на клавішу *Variable*.

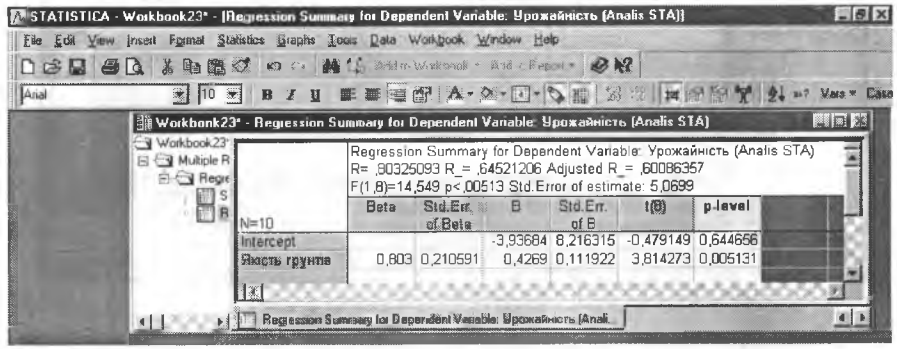
5. У вікні, яке з'явиться, в лівій частині таблиці оберіть результативний показник (*Dependent var. — залежна змінна*), в правій — факторний показник (*Independent var. — незалежна змінна*). В нашому прикладі рівень урожайності (результативний показник) залежить від якості ґрунтів (факторний показник). Двічі натисніть клавішу *OK*.



На екрані з'явиться вікно *Multiple Regression Results.*



6. При подальшому натисканні на кнопку *Summery Regression: results* ми отримаємо електронну таблицю, з результатами регресійного аналізу залежності урожайності озимої пшениці від якості ґрунтів.



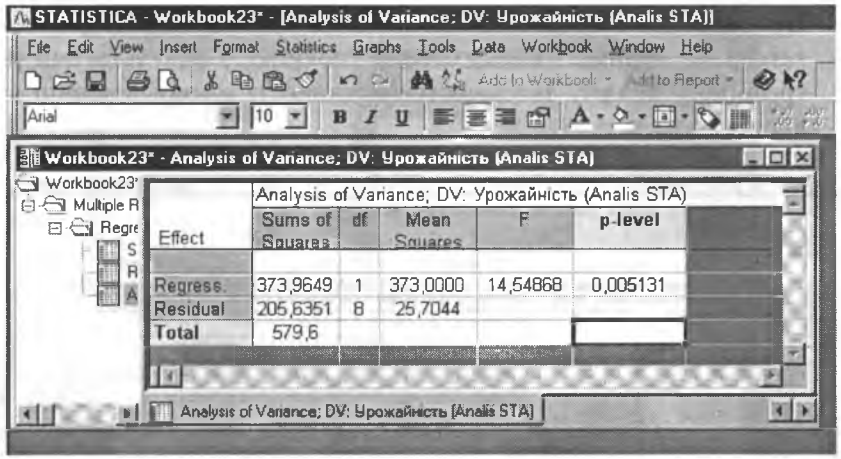
Так, рівняння регресії має такий вигляд:

$$y_x = -3{,}93684 + 0{,}4269x\,.$$

Коефіцієнт регресії показує, що при збільшенні якості ґрунтів на 1 бал, приріст урожайності озимої пшениці становить 0,4269 ц/га.

Зв'язок між досліджуваними показниками тісний, оскільки лінійний коефіцієнт парної кореляції становить $r_{y_x} = 0{,}803$. Коефіцієнт детерміиації 0,645 показує, що 64.5 % загальної варіації результативної ознаки (урожайності озимої пшениці) зумовлені дією включеного в кореляційну модель фактора (якістю ґрунтів). На долю не врахованих факторів приходиться 35,5 %.

7. Для того, щоб отримати результати дисперсійного аналізу, необхідно натиснути на клавішу *ANOVA (Overall goodness of fit)* (див. п. 5). Електронна таблиця, в якій наведено результати аналізу варіацій має вигляд:



Отже, розкладання загальної варіації урожайності озимої пшениці на внутрішньо- і міжгрупову, дало такі результати:

- загальна варіація $w_3 = 579{,}6$;
- міжгрупова (систематична) варіація $w_M = 373{,}9649$;
- внутрішньогрупова (випадкова) варіація $w_B = 205{,}6351$.

Фактичне значення F – критерію для якості ґрунтів за електронною таблицею становить 14,54868. Ми можемо перевірити правильність розрахунків. Визначимо фактичне значення F – критерію як відношення факторної дисперсії до залишкової:

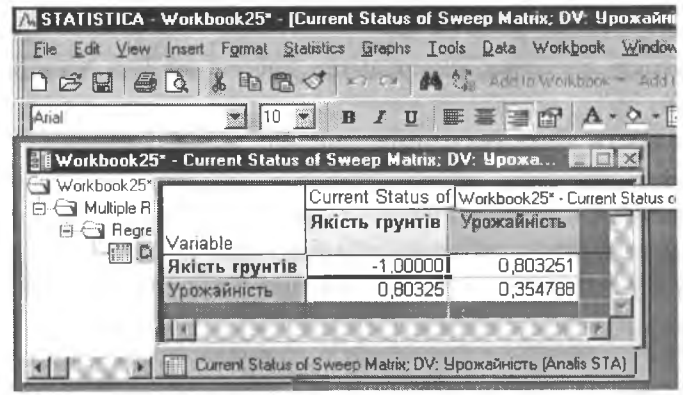
$$F = \frac{373}{25{,}7044} = 14{,}54868 .$$

Табличне значення F – критерію за таблицею «Значення F» при імовірності 0,95» становить 5,12. Фактичне відношення між порівнюваними дисперсіями значно перевищує табличне:

$$14{,}54868 > 5{,}12 .$$

Це дає змогу зробити висновок, що різниця між урожайністю у варіантах досліду не пов'язана з випадковим варіюванням, а є істотною і зумовлена впливом якості ґрунтів.

8. При натисканні на клавішу *Current sweep matrix* (пункт 5) на екрані з'явиться матриця парних лінійних коефіцієнтів кореляції.



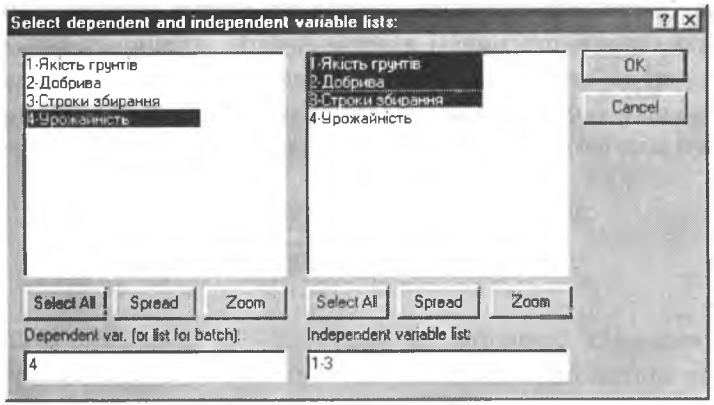
Так, частковий лінійний коефіцієнт парної кореляції між урожайністю озимої пшениці і якістю ґрунтів становить 0,80325. Це значить, що зв'язок між досліджуваними ознаками, по-перше, прямий (оскільки коефіцієнт має додатне значення), по-друге, — тісний (оскільки абсолютне значення коефіцієнта більше 0,7).

8.6 Множинна кореляція

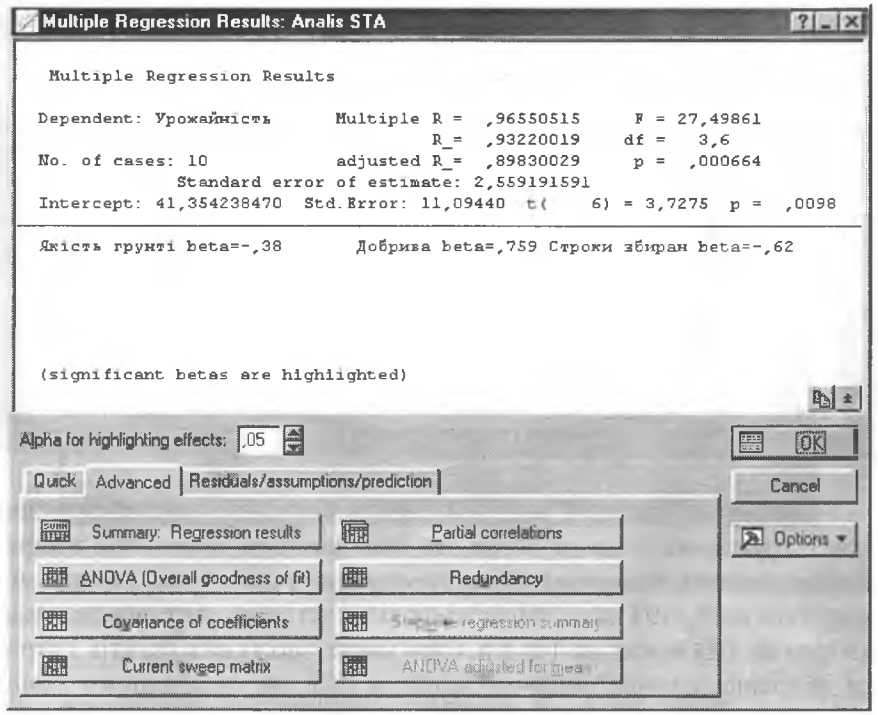
Порядок визначення параметрів рівняння множинної регресії, множинного (сукупного) коефіцієнта кореляції, детермінації, коефіцієнтів парної кореляції розглянемо на прикладі залежності урожайності озимої пшениці одночасно від трьох факторів:

- якості ґрунтів, балів (x_1);
- кількості внесених мінеральних добрив на 1 га посівів, кг д.р. (x_2);
- тривалість збирання озимої пшениці, днів (x_3).

1. У випадаючому меню *Statistics* оберіть модуль *Multiple Regression* в режимі *Advanced* натиснути на клавішу *Variable*. В лівій частині таблиці оберіть результативний показник (*Dependent var. — залежна змінна*) — урожайність озимої пшениці, в правій — факторні показники (*Independent var. — незалежна змінна*) — якість грунтів, добрива, строки збирання. Двічі натисніть клавішу *OK*.



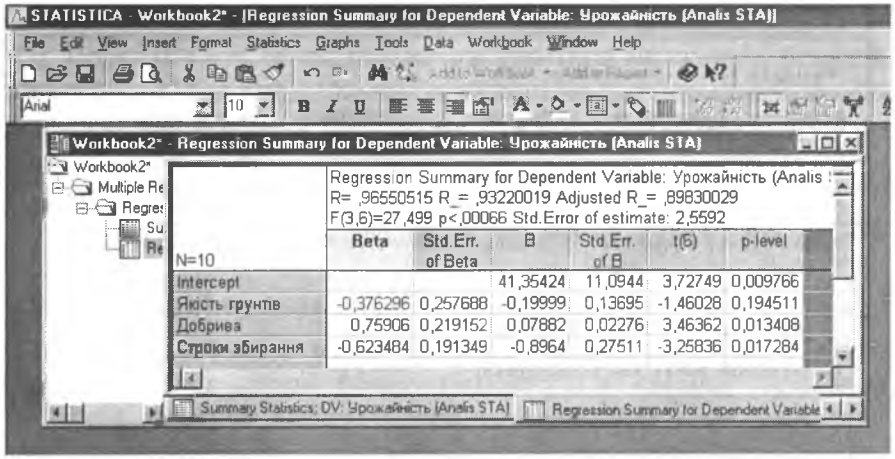
2. На екрані з'явиться вікно *Multiple Regression Results*.



3. Для отримання параметрів рівняння регресії, вкажіть курсором на клавішу *Summery Regression: results*. Ми отримаємо електронну таблицю, з результатами множинного регресійного аналізу залежності урожайності озимої пшениці від якості ґрунтів, кількості внесених мінеральних добрив на 1 га посівів і строків збирання.

Рівняння множинної регресії має такий вигляд:

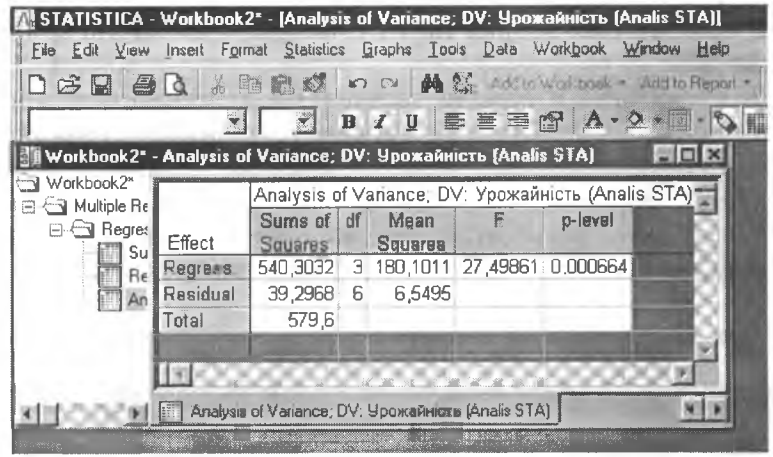
$$y_{x_1,x_2,x_3} = 41.35424 - 0{,}1999x_1 + 0{,}07882x_2 - 0{,}8964x_3$$



Коефіцієнт регресії $a_1 = -0.1999$ показує, що при збільшенні якості ґрунтів на 1 бал за умови, що інші фактори, включенні в кореляційну модель залишаються на середньому рівні, урожайність зменшується на 0,1999 ц/га. Збільшення кількості внесених мінеральних добрив на 1 га посіві на 1 ц д.р. (при умові, що якість ґрунтів і строки збирання озимої пшениці залишаються на фіксованому рівні) сприяє зростанню урожайності на 0,07882 ц/га. Внаслідок збільшення тривалості збирання на 1 день за умови середнього рівня інших факторів, урожайність озимої пшениці зменшується на 0,8964 ц/га.

Над таблицею, в другому рядку, ми бачимо значення множинного коефіцієнта кореляції і детермінації. Коефіцієнт сукупної кореляції становить $R = 0.966$, що свідчить про тісний, суттєвий зв'язок між досліджуваними ознаками. Значення коефіцієнта множинної детермінації $R^2 = 0.932$ показує, що на долю систематичної варіації урожайності озимої пшениці, зумовленою дією факторних показників, включених до кореляційної моделі, приходиться 93,2 %. На долю не врахованих в досліді факторів залишається 6,8 %.

4. Результати дисперсійного аналізу ми отримаємо, якщо натиснемо на клавішу *ANOVA (Overall goodness of fit)*.



Дані електронної таблиці показують такі результати дисперсійного аналізу:

- загальна варіація $w_3 = 579.6$;
- міжгрупова (систематична) варіація $w_M = 540{,}3032$;
- внутрішньогрупова (випадкова) варіація $w_B = 39{,}2968$.

Фактичне значення F – критерію для якості ґрунтів за електронною таблицею становить 27,49861 ($F = \dfrac{180{,}1011}{6{,}5495} = 27{,}49861$).

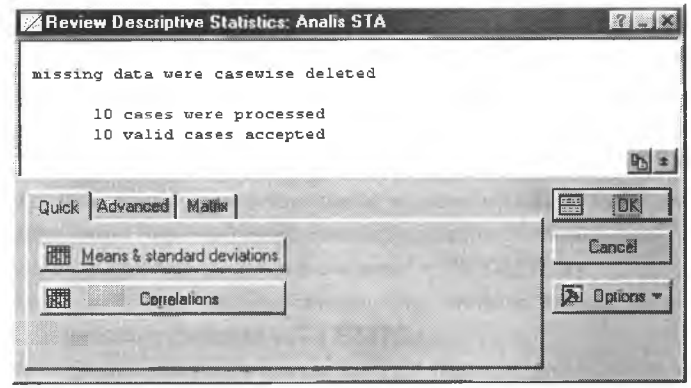
Табличне значення F – критерію за таблицею «Значення F» при імовірності 0,95» становить 4,76. Фактичне відношення між порівнюваними дисперсіями значно перевищує табличне:

$$27{,}49861 > 4{,}76.$$

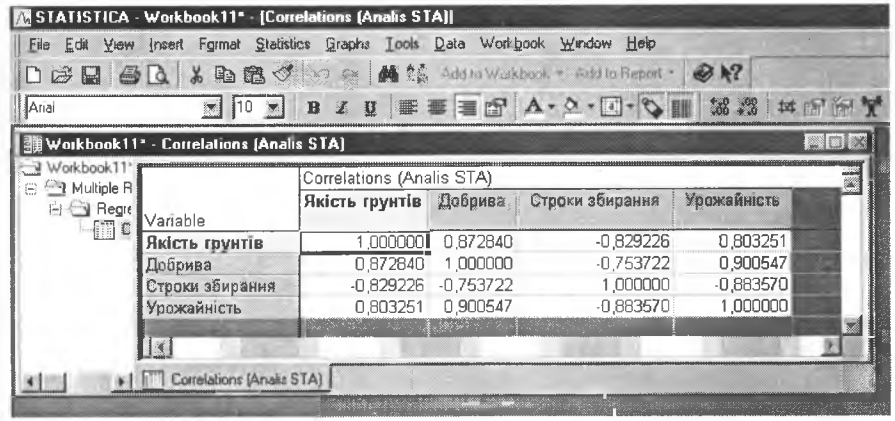
Це дає змогу зробити висновок, що різниця між урожайністю у варіантах досліду не пов'язана з випадковим варіюванням, а є істотною і зумовлена впливом якості ґрунтів, кількості внесених мінеральних добрив на 1 га посівів, тривалістю збирання урожаю.

5. Для того, щоб отримати часткові парні коефіцієнти кореляції, які покажуть напрям і тісноту залежності урожайності від кожного окремо взятого фактора, в модулі *Multiple Regression* в режимі *Advanced* вкажіть курсором на клавішу *Review descriptive statistics, correlation matrix*. Натисніть клавішу *OK*.

6. На екрані з'явиться робоче вікно *Review Descriptive Statistics*, в якому необхідно обрати клавішу *Correlation*.



7. На екрані з'явиться матриця парних коефіцієнтів кореляції.



У нашому прикладі коефіцієнти парної кореляції дорівнюватимуть:

- між урожайністю та якістю ґрунтів $r_{yx_1} = 0{,}803251$;
- між урожайністю та кількістю внесених добрив $r_{yx_2} = 0{,}900547$;
- між урожайністю та строками збирання $r_{x_1x_2} = -0{,}88357$.

Обчислені парні коефідієнти кореляції показують, що урожайність озимої пшениці перебуває в прямому тісному зв'язку з якістю ґрунтів і кількістю внесених на 1 га посівів мінеральних добрив. Залежність урожайності від термінів збирання урожаю носить тісний обернений характер.

Контрольні питання

1. *Дайте характеристику інтегрованої системи аналізу й управління даними «STATISTICA».*
2. *Порядок настройки системи «STATISTICA».*
3. *Розкрийте порядок введення даних в систему «STATISTICA».*
4. *Найпростіші описові статистики і порядок їх обчислення.*
5. *Порядок побудови графіків в системі «STATISTICA».*
6. *Порядок побудови лінійної регресійної моделі, перевірки параметрів на значимість в системі «STATISTICA».*
7. *Проінтерпретувати таблицю ANOVA-дисперсійного аналізу.*
8. *Порядок проведення багатофакторного регресійного аналізу в системі «STATISTICA».*

ОСНОВНІ КАТЕГОРІЇ І ТЕРМІНИ

Автокореляція — кореляцію (залежність) значень y_t і $y_t + h$, де t — поточний момент часу, h — період. А. наявна, коли теперішні та лагові значення випадкової величини корелюють. А. впливає на незалежність спостережень і спричинює викривлення результатів аналізу.

Авторегресія — регресія, що враховує вплив попередніх рівнів ряду на наступні.

Альтернативна гіпотеза — гіпотеза, яка протиставляється нульовій гіпотезі і заперечує її.

Альтернативні ознаки — ознаки, які мають одні одиниці сукупності, а інші не мають (прибуткове чи збиткове підприємство). Є лише два варіанти, причому один з них виключає інший, наявність ознаки позначають через 1, а її відсутність — через 0.

Аналіз — метод наукового дослідження об'єкта шляхом розгляду його окремих аспектів, складових частин.

Аномальні спостереження — спостереження, характерні для нестабільних явищ і процесів. Виявляються і коригуються за допомогою методів Ірвіна і q - статистики.

Атрибутивний ряд розподілу — ряд, побудований за якісною ознакою.

Атрибутивні ознаки — ознаки, які не мають кількісного виразу і реєструються у вигляді текстового запису (наприклад, марки машин, професії працівників).

Багатоступінчаста вибірка — типова вибірка, яку проводять кількома стадіями (ступенями). При цьому кожна стадія має свою одиницю відбору. При багатоступінчастому відборі з генеральної сукупності

спочатку відбирають укрупнені групи, потім — більш дрібні і так доти, доки не будуть відібрані ті одиниці, що піддаються обстеженню.

Багатофазна вибірка — вибірка, яка припускає збереження однієї й тієї ж одиниці добору на всіх етапах його проведення. При цьому відібрані на кожній стадії одиниці піддаються обстеженню. На кожній наступній стадії добору програма обстеження розширюється.

Безповторна власне випадкова вибірка — вибірка, при якій кожна раніше відібрана одиниця не повертається в генеральну сукупність і в дальшому відборі не бере участі.

β-коефіцієнти показують, на скільки середніх квадратичних відхилень змінюється результативний показник при зміні відповідного фактора на одне значення середньоквадратичного відхилення. Характеризують вплив окремих факторів на результативну ознаку.

Бімодальний розподіл — розподіл, коли два варіанти мають найбільші й однакові частоти, тобто ряд має дві моди. Бімодальний розподіл вказує на якісну неоднорідність сукупності за досліджуваною ознакою.

Біссеріальний коефіцієнт кореляції — оцінювання зв'язку між якісною альтернативною та кількісною варіативною ознаками.

Варіанти — окремі значення ознаки, які вона приймає у варіаційному ряді.

Варіаційний ряд розподілу — ряд, побудований за кількісною ознакою.

Варіація — коливання значень ознаки. Також: коливання, різноманіття, змінюваність величини ознаки (зміна розміру ознаки) в окремих одиниць сукупності.

Варіативна ознака — ознака, яка має в межах сукупності різні значення.

Величина інтервалу — різниця між верхньою і нижньою межами інтервалу.

Вибіркова сукупність — частина генеральної сукупності, яку вибірково обстежуватимуть, сукупність відібраних для обстеження одиниць.

Вибіркова частка — питома вага одиниць, що володіють даною ознакою у вибірковій сукупності. Розходження між вибірковою часткою і середнім значенням ознаки у вибірці (вибіркової середньої) визначають особливості обчислення необхідного обсягу, помилок вибірки, інтервалів довіри та ін.

Вибіркове (репрезентативне) спостереження — найбільш поширений вид несуцільного спостереження, при якому закономірності й характеристики, властиві якісь генеральній сукупності, визначають дослідженням деякої її частини. Його основою є випадковий відбір одиниць для обстеження, що гарантує незалежність результатів вибірки від волі осіб, які його проводять.

Вибірковий метод — сукупність математичних способів і обґрунтувань, які використовують при застосуванні вибіркового спостереження.

Викиди — це нетипові спостереження, які дуже вирізняються зпосеред інших.

Відкриті інтервали — інтервали, в яких невідомі мінімальні та максимальні значення. Відкритими можуть бути перший або останній інтервали.

Відносна щільність розподілу — відношення частості до величини того ж інтервалу в інтервальних рядах розподілу з нерівними інтервалами.

Відносні показники варіації — коефіцієнти осцилляції, варіації, відносне лінійне відхилення.

Відстань Чебишева — це відстань, яка використовується при необхідності визначення двох об'єктів як «різних», якщо вони відрізняються за будь-якою однією координатою (будь-яким одним виміром): *відстань* $(x, y) = \max|x_i - y_i|$.

Власне випадкова вибірка — вибірка, при якій кожна одиниця з генеральної сукупності відбирається у вибірку випадково, ненавмисно. При цьому генеральна сукупність не розподіляється на складові частини. Відбір одиниць звичайно проводиться жеребкуванням.

Внутрішньогрупова варіація — варіація, яка характеризує коливання, зумовлене факторами, не врахованими під час групування. Вона дорівнює сумі квадратів відхилень індивідуальних значень ознаки від групових середніх в усіх групах.

Внутрішньогрупова дисперсія — відбиває випадкову варіацію, тобто частину варіації, що відбувається під впливом неврахованих факторів і не залежить від факторної ознаки.

Генеральна сукупність — загальна сукупність одиниць, з якої проводять відбір частини одиниць.

Гранична помилка — це t разів узята середня помилка.

Густина розподілу — відношення частоти до величини інтервалу в інтервальних рядах розподілу з нерівними інтервалами.

Групувальна ознака — ознака, за якою проводиться розбивка одиниць сукупності на окремі групи.

Групування — розчленовування безлічі одиниць досліджуваної сукупності на групи за певними, суттєвими для них ознакам.

Деревовидна кластеризація — об'єднання об'єктів дослідження в досить великі кластери, використовуючи деяку міру схожості або відстань між об'єктами.

Децилі — додаткові статистичні характеристики рядів розподілу, які поділяють ранжируваний ряд розподілу на 10 рівних частин.

Дискретні варіаційні ряди розподілу — ряди, в яких варіанти виражено цілими числами (розподіл домогосподарств України за кількістю дітей віком до 18 років).

Дисперсія — середній квадрат відхилень індивідуальних значень ознаки від їхньої середньої величини.

Евклідова відстань — найбільш загальний тип відстані. Є геометричною відстанню в багатомірному просторі:

$$\text{відстань } (x, y) = \rho(x, y) = \left\{\sum_i (x_i - y_i)^2\right\}^{\frac{1}{2}}.$$

Економіко-статистичний аналіз — розробка методики, заснованої на широкому застосуванні традиційних статистичних і математико-статистичних методів, з метою контролю адекватного відбиття досліджуваних явищ і процесів.

Економічна інтерпретація моделі — висновки на основі розрахунку та аналізу парних коефіцієнтів еластичності, парних і множинного коефіцієнтів детермінації, β - коефіцієнта.

Експлікація — опис змісту графіка за допомогою слів (словесний опис).

Екстраполяція — знаходження рівнів за межами досліджуваного ряду, тобто продовження ряду на основі виявленої закономірності зміни рівнів у досліджуваний період.

Емпіричне кореляційне відношення — корінь квадратний з емпіричного коефіцієнта детермінації.

Емпіричний коефіцієнт детермінації — частка міжгрупової дисперсії в загальній дисперсії.

Ентропія — міра невизначеності даних спостереження, яка може мати різні результати. Залежить від кількості градацій ознаки та ймовірності кожної з них.

Загальна варіація ознаки — варіація, яка характеризує вплив усіх умов на варіацію ознаки. Її обчислюють як суму квадратів відхилень індивідуальних значень ознаки x від загальної середньої $\bar{x}_з$.

Загальна середня — середня, що показує типовий розмір ознаки якісно однорідної сукупності в цілому.

Загальні методи кластерного аналізу — об'єднання (деревоподібна кластеризація), двовходове об'єднання, метод *K* середніх.

Закон великих чисел — зі збільшенням кількості спостережень вплив випадкових причин, що визначають величину ознаки окремих одиниць сукупності, в цілому взаємно погашається й у зведених характеристиках виражається дія основних причин, тобто визначається закономірність.

Закон розподілу — це співвідношення між можливими значеннями ознаки величин і відповідними ймовірностями.

Закономірність — повторюваність, послідовність і порядок змін у явищах.

Закономірності розподілу — закономірності зміни частот у варіаційних рядах.

Закриті інтервали — інтервали, в яких відомі мінімальні і максимальні значення ознаки (у яких позначено обидві межі).

Залежні вибірки — вибірки, коли спостереження в одній вибірці якоюсь мірою пов'язано зі спостереженнями у другій.

Зважений попарний центроїдний метод (медіана) — при обчисленнях використовуються ваги для врахування різниці між розмірами кластерів (тобто числом об'єктів у них).

Зважене попарне арифметичне середнє — аналогічний методу незваженого попарного середнього за винятком того, що при обчисленнях розмір відповідних кластерів, використовується як ваговий коефіцієнт.

Зведення — комплекс послідовних операцій з узагальнення конкретних одиничних факторів для виявлення типових рис і закономірностей, властивих досліджуваному явищу в цілому.

Звітна одиниця — суб'єкт, від якого надходять дані про одиницю спостереження.

Зворотний зв'язок — зменшення чи збільшення значення результативної ознаки із збільшенням чи зменшенням значень факторної ознаки.

Імовірність — міра об'єктивної можливості здійснення певних подій. Кількісно імовірність виражають відношенням кількості сприятливих наслідків до кількості можливих наслідків.

Індекс кореляції (кореляційне відношення) — показник тісноти зв'язку між факторною і результативною ознаками при криволінійній формі зв'язку, який показує, яку частку в загальному середньоквадратичному відхиленні результативної ознаки становить середньоквадратичне відхилення факторної ознаки. Змінюється в межах від 0 до +1, тобто завжди є додатною величиною.

Інтервал групування — різниця між максимальними і мінімальними значеннями ознаки в кожній групі.

Інтервальна структура — структура, що характеризує склад соціально-економічних явищ за певні періоди часу (дні, тижні, місяці, квартали, роки).

Інтервальний ряд динаміки — ряд числових значень статистичного показника, що характеризує розміри досліджуваного явища за певні проміжки (періоди, інтервали) часу.

Інтервальні ряди розподілу — ряди, в яких варіанти виражені у вигляді інтервалів. Інтервали в рядах розподілу можуть бути рівними і нерівними.

Інтерполяція — наближений розрахунок рівнів, що лежать усередині ряду динаміки, але з якихось причин невідомих.

Квадрат евклідової відстані — використовується у випадках, коли виникає необхідність надати більші ваги віддаленим один від одного об'єктам: *відстань* $(x, y) = \sum_i (x_i - y_i)^2$.

Квадрат середньої помилки вибіркового коефіцієнта регресії — відношення скоригованої залишкової дисперсії до суми квадратів відхилень незалежної змінної: $m^2 = \dfrac{S^2}{\sum (x - \bar{x})^2}$.

Квартилі — додаткові статистичні характеристики рядів розподілу, що поділяють ряд розподілу за сумою частот на 4 рівні частини.

Кількісні ознаки — ознаки, що реєструються числом. Групувальна ознака може бути виражена числами по-різному. Одні ознаки виражаються тільки цілими числами. Така ознака зветься дискретною, або перервною. Інші ознаки можуть позначатися цілими і дробовими числами. Ці зміни ознаки називають безперервними.

Класифікація — різновид типологічних групувань, систематизований, заздалегідь встановлений поділ явищ і об'єктів на групи, класи, розряди, категорії за якими проводиться зведення даних. Основою класифікації, як правило, є якісна ознака.

Кластер — однорідні групи, отримані в результаті розподілу сукупності об'єктів, кожний з яких характеризується набором k - ознак.

Кластерний аналіз — метод знаходження кластерів.

Коефіцієнт асоціації та контингенції — визначення тісноти зв'язку між двома альтернативними ознаками, кожна з яких складається тільки з двох груп.

Коефіцієнт варіації — відсоткове відношення середнього квадратичного відхилення до середньої величини ознаки.

Коефіцієнт взаємного узгодження — застосовують для вимірювання тісноти зв'язку між двома атрибутивними ознаками, коли їх варіювання утворює кілька (більше двох) груп.

Коефіцієнт. детермінації — показує, на скільки відсотків варіація результативної ознаки пояснюється варіацією i -ї ознаки (парний К. д.) чи усіх факторних ознак, що ввійшли в модель (множинний К. д.).

Коефіцієнт еластичності — показує, на скількох відсотків у середньому зміниться значення результативної ознаки за зміни факторної ознаки на 1 %.

Коефіцієнт конкордації — визначення тісноти зв'язку між довільною кількістю ранжируваних ознак.

Коефіцієнт кореляції рангів — визначення тісноти зв'язку між двома кількісними або якісними ознаками після попереднього ранжирування їх в порядку зростання або спадання.

Коефіцієнт множинної (сукупної) детермінації — показує, яка частка варіації досліджуваного результативного показника зумовлена впливом факторів, включених у рівняння множинної регресії. Він може мати значення від 0 до +1.

Коефіцієнт множинної (сукупної) кореляції — основний показник тісноти зв'язку за множинної кореляції, який може мати значення від 0 до +1.

Коефіцієнт осцилляції — відсоткове відношення розмаху варіації до середньої величини ознаки.

Коефіцієнт регресії — показує, наскільки в середньому змінюється значення результативної ознаки зі зміною факторної на одиницю власного виміру.

Коефіцієнт Фехнера — застосовують для оцінки тісноти зв'язку порівнянням знаків відхилень варіантів факторної та результативної ознак від середніх.

Коефіцієнти асоціації та контингенції — використовуються для визначення тісноти зв'язку двох якісних ознак, кожна з яких складається тільки з двох груп.

Коефіцієнти еластичності — показують, на скільки відсотків змінюється результативна ознака зі зміною факторної ознаки на 1 %.

Коліпеарність (мультиколінеарність) — кореляційний зв'язок між факторами в рівнянні множинної регресії. Чим вища колінеарність, тим менш надійними будуть показники впливу окремих факторів.

Кореляційна залежність — це функціональне співвідношення лише між середніми значеннями досліджуваних ознак.

Кореляційне відношення — показує зв'язок між двома ознаками.

Кореляційне поле — спосіб графічного зображення взаємозалежності статистичних показників. Дає наочне уявлення про наявність зв'язку між досліджуваними ознаками.

Кореляційний аналіз — це метод визначення і кількісної оцінки взаємозалежностей між статистичними ознаками, що характеризують окремі соціально-економічні явища і процеси.

Кореляційний зв'язок — різновид стохастичного зв'язку, що виявляється в зміні середніх умовних розподілів. При К. з. немає суворої відповідності між значеннями залежних ознак: кожному певному значенню аргументу (факторної ознаки) відповідає кілька різних значень функції (результативної ознаки).

Кореляція — статистична залежність між випадковими величинами, що не має суто функціонального характеру, за якої зміна однієї з випадкових величин приводить до зміни математичного очікування іншої.

Крива розподілу — графічне зображення у вигляді безперервної лінії зміни частот у варіаційному ряді, функціонально зв'язаного зі зміною варіант.

Криволінійний кореляційний зв'язок — зв'язок, який характеризується тим, що однаковим змінам середніх значень факторної ознаки відповідають різні зміни середніх значень результативної ознаки. Аналітично криволінійний зв'язок визначають за рівнянням кривої лінії.

Критерії узгодженості — особливі статистичні показники, що характеризують відповідність емпіричного й теоретичного розподілів. Відомі критерії згоди Д. Пирсона, В. Романовського, А. Н. Колмогорова, Б. З. Ястремського.

Критерій λ Колмогорова — застосовують для наближеної оцінки імовірності розбіжностей між фактичними й теоретичними розподілами. При цьому не потрібно визначати кількість ступенів свободи варіації, оскільки граничні значення критерію не залежать від кількості спостережень і є стандартними.

Критерій t нормального розподілу — це теоретичне нормоване відхилення для великих виборок. За законом нормального розподілу варіація індивідуальних значень досліджуваної ознаки перебуває в межах $\bar{x} \pm 3\sigma$ (правило трьох сигм). Числове значення цього критерію залежить від рівня імовірності. Його визначають за спеціальними таблицями «Значення інтеграла імовірностей».

Критерій t-Стьюдента використовують для перевірки статистичних гіпотез стосовно середніх при малій вибірці (п <. 20). Його застосовують, визначаючи надійні інтервали, інтервально оцінюючи параметри генеральної сукупності. Числове значення критерію залежить від кількості ступенів свободи варіації та рівня ймовірності.

Критерій F-Фішера-Снедекора використовують у оцінці співвідношення дисперсій при малих вибірках, а також суті ступеня варіації ознак і надійності взаємозв'язку між факторами.

Критерій Вілкоксона застосовують для перевірки однорідності розподілів двох генеральних сукупностей.

Критерій χ^2 Пірсона — використовують тоді, коли потрібно визначити ступінь відмінності фактичного розподілу частот від теоретичного. Крім того, його застосовують для оцінки однорідності розподілів, а також як критерій незалежності в розподілі об'єктів сукупності за градаціями досліджуваної ознаки.

Критична область — це ті значення критерію, при яких нульова гіпотеза відхиляється.

Критична оцінка вихідних даних — повнота, якість і вірогідність відповідності емпіричного матеріалу цілям і завданням дослідження.

Критичні точки — точки, які відокремлюють критичну галузь від галузі допустимих значень.

Лаг — проміжок часу відставання одного явища від іншого, пов'язаного з ним.

Лінії регресії — лінії, побудовані на основі рівнянь регресії.

Лінійний зв'язок — статистичний зв'язок між явищами, виражений рівнянням прямої лінії.

Лінійний коефіцієнт варіації — відношення середнього лінійного відхилення до середнього значення ознаки.

Лінійний коефіцієнт парної кореляції — кількісний показник тісноти прямолінійного зв'язку результату з одним фактором При парній залежності коефіцієнт кореляції коливається від 0 до +1 за прямого зв'язку і від 0 до –1 — за зворотного зв'язку. Якщо $r < 0.3$, зв'язку немає, якщо $r = 0.3 - 0.5$ — зв'язок слабкий, якщо $r = 0.5 - 0.7$ — зв'язок середній і якщо $r > 0.7$ — зв'язок тісний.

Мала вибірка — вибіркове спостереження, чисельність одиниць якого не перевищує 20. При малій вибірці діє особливий закон розподілу. Величина ймовірної помилки залежить як від коефіцієнта довіри t, так і від обсягу вибірки за випадку, якщо гранична помилка не перевищить t-кратну середню помилку в малих вибірках.

Манхеттенська відстань (відстань міських кварталів) — є середньою різницею між координатами. У більшості випадків призводить до таких же результатів, як і для відстані Евкліда: $відстань\ (x, y) = \sum_i |x_i - y_i|$.

Матеріали спостереження — це первинна статистична інформація, яка є основою для одержання узагальнювальних характеристик.

Матриця — прямокутна таблиця числової інформації, що складається з m рядків і n стовпчиків.

Медіана (структурна або розподільна середня) — значення ознаки в одиниці сукупності, що займає середнє положення в ранжированому ряду розподілу. Вона є центром розподілу сукупності і ділить її на дві рівні за кількістю частини.

Метод Варда — заснований на використанні методів дисперсійного аналізу для оцінки відстаней між кластерами.

Метод головних компонент — засіб зниження розмірності. Використовується і для проведення класифікацій. Суть методу полягає у виділенні лінійних комбінацій вихідних факторних ознак, які мають максимально можливу дисперсію. При цьому, перша головна компонента володіє максимальною дисперсією і є нормованою лінійною комбінацією всіх можливих вихідних ознак, а друга — враховує максимальне значення дисперсії, що залишилась, і кореляційно не пов'язана з першою компонентою.

Метод повного зв'язку — заснований на використанні відстані між одиницями кластерів, які найбільш віддалені від інших пар об'єктів і один від одного.

Механічна вибірка — різновид випадкової вибірки, коли одиниці для вибіркового спостереження відбирають не жеребкуванням, а механічно через відповідний інтервал. Для цього всі одиниці генеральної сукупності розподіляють у певному порядку, але так, щоб порядок не був пов'язаний із розміром досліджуваної ознаки. М. в. завжди безповторна.

Механічне згладжування — метод знаходження плавних рівнів ряду динаміки шляхом використання ковзних середніх. Розрізняють метод незважених і зважених ковзних середніх.

Міжгрупова варіація ($w_м$) — варіація, що характеризує систематичну варіацію, зумовлену факторною ознакою, за якою проведене групування.

Вона дорівнює сумі квадратів відхилень групових середніх $\overline{x_{гр}}$ від загальної середньої $x_з$.

Міжгрупова дисперсія характеризує систематичну варіацію, тобто розходження у величині досліджуваної ознаки, що виникають під дією ознаки — фактора, покладеного в основу групування.

Міри відстані: Евклідова відстань; квадрат евклідової відстані; Хемінгова відстань; Манхеттенська відстань; відстань Чебишева; степенева відстань.

Множинна кореляція — кореляція, за допомогою якої вивчається вплив на результативну ознаку двох і більше взаємопов'язаних факторних ознак.

Множинна регресія — модель зв'язку трьох і більш ознак.

Множинний коефіцієнт кореляції — відбиває зв'язок між результативною і декількома факторними ознаками.

Мода (структурна або розподільна середня) — значення ознаки, яке найчастіше повторюється в досліджуваній сукупності. Це варіант, який має найбільшу частоту.

Мода і медіана — структурні середні.

Момент другого порядку — середній квадрат зі значень $\left(\frac{x-a}{h}\right)^2$.

Момент першого порядку — середня величина зі значень $\frac{x-a}{h}$.

Моментний ряд динаміки — ряд числових значень статистичного показника, що характеризує розміри досліджуваного явища на певні дати, моменти часу.

Мультиколінеарність — наявність тісної залежності між факторними ознаками.

Нагромаджені (кумулятивні) частоти (частості) — це сума частот (частостей) варіантів від мінімального значення ознаки до даного значення.

Незалежні вибірки — вибірки, коли кожному спостереженню в одній вибірці не можна протиставити спостереження другої вибірки (варіанти яких змінюються незалежно один від одного).

Незважений попарний центроїдний метод — відстань між двома кластерами визначається як відстань між їх центрами ваги.

Незважене попарне арифметичне середнє — відстань між двома різними кластерами визначається як середня відстань між усіма парами об'єктів у них.

Нелінійний зв'язок — статистичний зв'язок між соціально-економічними явищами, аналітично виражений рівнянням кривої лінії (параболи, гіперболи і т. ін.).

Непараметричні критерії — статистичні критерії, використання яких не пов'язане зі знанням закону розподілу випадкової величини. Їх можна використовувати і тоді, коли досліджуваний розподіл відрізняється від нормального. До непараметричних належать критерії Колмогорова, Вілкоксона, Уайта тощо.

Несуцільне статистичне спостереження — охоплює лише певну частину одиниць досліджуваного об'єкта для одержання узагальнювальних характеристик об'єкта в цілому (обстеження домогосподарств охоплюють не всі домашні господарства). Залежно від характеру відбору одиниць для спостереження розрізняють такі види н. с. с.: вибіркове; основного масиву; анкетне; монографічне.

Нормальний розподіл — це симетричний розподіл, в якому максимуми значень випадкової величини концентруються навколо середньої величини.

Нормоване відхилення — відношення помилки вибірки до середньої помилки.

Об'єкт статистичного спостереження — сукупність суспільних явищ і процесів, про які слід зібрати статистичні відомості. О. с. с. органів державної статистики є галузі національної економіки, ресурси виробництва, населення тощо. В окремих підприємствах О. с. с. можуть бути наявність та результати використання земельних, трудових і матеріальних ресурсів.

Обернений кореляційний зв'язок — зв'язок, за якого зі збільшенням факторної ознаки результативна ознака зменшується чи, навпаки, зі зменшенням факторної ознаки результативна зростає.

Об'єднання (деревовидна кластеризація) — об'єднання об'єктів дослідження в досить великі кластери, використовуючи деяку міру схожості або відстань між об'єктами.

Об'єкт спостереження — статистична сукупність, у якій виникають досліджувані соціально-економічні явища і процеси.

Область допустимих значень — це ті значення критерію, при яких нульова гіпотеза приймається.

Обсяг сукупності — кількість одиниць, які мають певну ознаку (кількість підприємств, працівників).

Обсяг явища — загальний розмір явищ (чисельність працівників, обсяг продукції тощо) для всієї сукупності.

Одиниці сукупності — окремі об'єкти (елементи, явища і т. ін.), які становлять статистичну сукупність. Вони характеризуються індивідуальними ознаками, на підставі яких статистика узагальнює дані сукупності в цілому.

Одиниця спостереження — складовий елемент об'єкта, що є носієм ознак, які підлягають реєстрації. Також: первинний елемент об'єкта дослідження, який є носієм ознак, що підлягають реєстрації.

Одиниця статистичної сукупності — кожен окремо взятий елемент даної множини, що володіє визначеними ознаками.

Одинокий зв'язок (метод «найближчого сусіда») — відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами в різних кластерах.

Ознака — загальна властивість, характерна риса чи інша особливість одиниць сукупності, за якими можна вести статистичне спостереження або виміряти їх. Основна відмітна риса, особливість досліджуваного явища чи процесу.

Опитування — спосіб збирання статистичних даних, при якому відповіді на запитання формуляра записують зі слів опитуваної особи. Розрізняють три способи опитування: усне опитування; самореєстрація; кореспондентський спосіб.

Основна (нульова) гіпотеза — гіпотеза, яка підлягає перевірці.

Основна тенденція (тренд) — досить плавна і стійка зміна рівня явища в часі, більш-менш вільна від випадкових коливань. Основну тенденцію можна подати або аналітично — у вигляді рівняння (моделі) тренда, або графічно.

Параметричні критерії — статистичні критерії, які ґрунтуються на припущенні, що розподіл досліджуваної ознаки в сукупності підпорядковується певному відомому закону, наприклад, законам: нормального розподілу, розподілу Стьюдента, Фішера і т. ін. До них належать критерії (t, F, χ^2). Особливістю цих критеріїв є те, що їх застосування потребує обчислення оцінок параметрів розподілу.

Парна кореляція — кореляційний зв'язок, при якому аналізують зв'язок між парою показників, один з яких факторний, другий — результативний.

Парна регресія — аналітичне вираження зв'язку двох ознак.

Парний коефіцієнт кореляції — показує ступінь тісноти зв'язку між двома ознаками при фіксованому значенні інших факторних ознак.

Парні коефіцієнти детермінації — добуток парних коефіцієнтів кореляції та — коефіцієнтів відповідних факторів. Використовують для розкладання загальної варіації результативного показника на включені в кореляційну модель фактори.

Парні коефіцієнти кореляції — показники для вимірювання тісноти зв'язку між двома досліджуваними ознаками без урахування їх взаємодії з іншими ознаками, включеними в багатофакторну кореляційну модель.

Перцентилі — значення ознаки, які ділять ряд розподілу на сто частин.

Повний зв'язок (метод «найбільш віддалених сусідів») — відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах (тобто «найбільш віддаленими сусідами»).

Повторна власне випадкова вибірка — вибірка, при якій кожна раніше відібрана одиниця знову повертається в генеральну сукупність і може знову брати участь у вибірці.

Поле графіка — частина площини, де розташовані графічні образи.

Помилка вибіркового спостереження — різниця між величиною параметра в генеральній сукупності та його величиною, обчисленою за результатами вибіркового спостереження.

Помилка репрезентативності — різниця між показниками вибіркової та генеральної сукупностей.

Помилка спостереження — розбіжність між розрахунковим і дійсним значеннями досліджуваних величин. Вони властиві всім несуцільним спостереженням, оскільки, хоч як правильно проводився відбір одиниць сукупності, узагальнювальні показники відібраної частини завжди будуть якоюсь мірою відрізнятися від відповідних показників усієї сукупності.

Помилки другого порядку — полягають у тому, що приймається нульова гіпотеза, хоч насправді правильна альтернативна гіпотеза.

Помилки першого порядку — полягають у тому, що відхиляється нульова гіпотеза, хоч насправді вона правильна.

Помилки статистичного спостереження — це розбіжності між розмірами якогось показника, що встановлені за допомогою спостереження, і справжніми його розмірами.

Правила об'єднання: одинокий зв'язок (метод «найближчого сусіда»), повний зв'язок (метод «найбільш віддалених сусідів»), незважене попарне арифметичне середнє, зважене попарне арифметичне середнє, незважений попарний цетроїдне усереднення, зважений попарний цетроїдний метод (медіана), метод Варда.

Правило складання (розкладання) варіації — загальна дисперсія ознаки дорівнює сумі міжгрупової та середньої з внутрішньогрупових дисперсій.

Проста гіпотеза — гіпотеза, яка стосується лише одного припущення.

Пряме перерахування і спосіб коефіцієнтів — методи поширення результатів вибіркового спостереження на генеральну сукупність.

Прямий зв'язок — із збільшенням чи зменшенням значень факторної ознаки збільшується чи зменшується значення результативної ознаки.

Прямий кореляційний зв'язок — зв'язок, при якому зміна факторної ознаки зумовлює зміну результативної ознаки в тому самому напрямі.

Прямолінійний кореляційний зв'язок — характеризується рівномірним зростанням або зменшенням результативної ознаки під впливом відповідної зміни факторної ознаки. Аналітичне його визначають за рівнянням прямої лінії.

Ранг — порядковий номер відповідної одиниці сукупності у ранжированому ряду. Також: порядковий номер значення ознаки, розміщеного в порядку зростання чи спадання величин.

Ранжирування — процедура упорядкування об'єктів вивчення, що виконується на основі переваги значень ознаки в порядку зростання чи спадання.

Регресійний аналіз — аналітичне вираження зв'язку, в якому зміна однієї величини — результативної ознаки — зумовлена впливом однієї чи декількох незалежних величин (факторів), а безліч всіх інших факторів, що також чинять вплив на залежну величину, приймається за постійні й середні значення.

Результативна ознака — ознака, яка змінюється під впливом факторної ознаки.

Репрезентативна вибірка — вибіркова сукупність, яка достатньо точно відображує генеральну сукупність.

Рівень значущості — імовірність припуститися помилки першого порядку. Р. з. становить ту мінімальну ймовірність, починаючи з якої можна визнати подію практично неможливою, тобто показує міру, з якою ми ризикуємо, відхиляючи нульову гіпотезу.

Рівень істотності — показує ймовірність, з якою гіпотеза, що перевіряється, може дати помилковий результат.

Рівняння регресії (кореляційне рівняння) — рівняння, за допомогою якого визначають статистичний зв'язок між величинами, що корелюють.

Розмах варіації — різниця між найбільшим і найменшим значеннями ознаки, що варіює.

Розмір кластерів — це число об'єктів, що містяться в кластерах.

Розподіл χ^2 (хі — квадрат) — це закон розподілу вибіркової дисперсії параметрів, які підпорядковуються закону нормального розподілу при малих вибірках.

Розподіл t-Стьюдента — це закон розподілу нормованого відхилення при малих вибірках ($n < 20$).

Ряд динаміки — ряд числових значень визначеного статистичного показника в послідовні моменти чи періоди часу.

Ряд розподілу — упорядкований розподіл одиниць сукупності на групи за певною варіативною ознакою.

Ряди динаміки — статистичні величини, які характеризують стан і зміну явищ у часі.

Середнє значення альтернативної ознаки — узагальнена характеристика варіативної ознаки сукупності, яка дорівнює частці одиниць сукупності, які мають певну ознаку.

Середнє квадратичне відхилення — мірило надійності середньої величини. Характеризує середнє коливання ознаки в сукупності, зумовлене індивідуальними особливостями одиниць сукупності. Обчислюють добуванням квадратного кореня з дисперсії.

Середнє лінійне відхилення — показник варіації, який становить середню з абсолютних відхилень усіх варіантів від середнього значення варіативної ознаки.

Середні величини — це узагальнювальні кількісні показники, що характеризують типові розміри варіативних ознак якісно однорідних сукупностей.

Середня арифметична — найбільш поширений вид середніх величин, який застосовують тоді, коли загальний обсяг варіативної ознаки для всієї сукупності становить суму індивідуальних значень усередненої ознаки. Визначають як відношення суми окремих значень ознаки до кількості одиниць сукупності.

Середня арифметична зважена — різновид середньої арифметичної, обчислювальний тоді, коли окремі значення усередненої ознаки повторюються в досліджуваній сукупності неоднакову кількість разів, а також для обчислення середньої зі середніх при різному обсязі сукупності. Зважування в цьому разі проводять за частотами, які показують, скільки разів повторюється певний варіант.

Середня арифметична проста — різновид середньої арифметичної, який застосовують тоді, коли відомі індивідуальні значення усередненої ознаки в кожної одиниці сукупності.

Середня помилка вибірки — середнє квадратичне відхилення вибіркових середніх від середньої генеральної сукупності.

Серійна (гніздова) вибірка — вибірка, при якій для спостереження відбирають не окремі одиниці генеральної сукупності, а серії (гнізда) таких одиниць. Відбір серій проводять власне випадковим або механічним способом. У відібраних серіях обчислюють усі без винятку одиниці.

Складна гіпотеза — гінотеза, що стосується двох і більше припущень і характеризує деяку область імовірних значень досліджуваного параметра.

Спосіб відліку від умовного нуля (способом моментів) — спосіб для обчислення середньої з варіаційного ряду з рівними інтервалами. При цьому одне зі значень варіативної ознаки, яке знаходиться в центрі розподілу або має найбільшу частоту, беруть за початок відліку. Потім від кожного усередненого значення інтервалу віднімають величину, взяту за початок відліку, і різницю ділять на величину інтервалу h. Добуті внаслідок цих перетворень нові варіанти утворюють у варіаційному ряду з рівними інтервалами прості числа — відхилення в інтервалах від умовного нуля.

Статистична гіпотеза — це припущення відносно параметрів або форми розподілу генеральної сукупності, яке можна перевірити на основі вибірки. У процесі перевірки статистичної гіпотези потрібно визначити, чи узгоджуються дані спостереження з висунутим припущенням. Унаслідок перевірки гіпотеза приймається або відхиляється.

Статистична оцінка параметра розподілу — наближене значення шуканої величини генеральної сукупності, встановлене на основі вибіркового спостереження.

Статистична сукупність — безліч одиниць, що володіють масовістю, однорідністю, визначеною цілісністю, взаємозалежністю станів окремих одиниць і наявністю варіації. Це сума об'єктів (подій, елементів, явищ тощо), які мають єдину якісну основу, але відрізняються певними ознаками. Найважливішою особливістю статистичної сукупності є однорідність, одноякісність її елементів.

Статистичний графік — креслення, на якому статистичні сукупності, які характеризуються певними показниками, описуються за допомогою умовних геометричних образів чи знаків.

Статистичний критерій — це, обчислений на основі фактичних спостережень оціночний показник, відповідно до якого приймають або відхиляють нульову гіпотезу.

Статистичний ряд розподілу — розподіл одиниць сукупності по групах за величиною варіативної ознаки. Такі ряди складаються з двох елементів: переліку груп і кількості одиниць, що входять до кожної групи. Вони характеризують склад сукупності за розміром досліджуваної ознаки.

Степенева відстань — застосовується у випадках, коли необхідно прогресивно збільшити або зменшити вагу, що належить до розмірності, для якої відповідні об'єкти дуже відрізняються:

$$\text{відстань } (x, y) = \left(\sum_i |x_i - y_i|^p\right)^{\frac{1}{r}}.$$

Стохастичний зв'язок — зв'язок, при якому кожному значенню ознаки відповідає певна множина значень ознаки y, які утворюють так званий умовний розподіл. Якщо умовні розподіли замінюють одним параметром — середнім значенням $\overline{y_s}$, то такий зв'язок називають кореляційним.

Ступінь свободи варіації — кількість варіантів, які можуть мати довільні значення, не змінюючи їх загальної характеристики (середньої).

Суцільне спостереження — одержання інформації про всі одиниці досліджуваної сукупності. Також: спостереження, при якому закономірності й характеристики визначають дослідженням всіх одиниць генеральної сукупності. Забезпечує найбільш повну інформацію про загальну кількість одиниць сукупності і дає вірогідні узагальнювальні статистичні характеристики явищ, що аналізуються.

Теоретична крива розподілу — крива, що виражає загальну закономірність даного типу розподілу в чистому вигляді, що виключає вплив випадкових факторів.

Типова (районована) вибірка — вибірка, при якій усю генеральну сукупність після попереднього аналізу розподіляють на однорідні, типові групи, райони, зони за певними ознаками. Потім з кожної групи у випадковому або механічному порядку відбирають певну кількість одиниць у вибіркову сукупність. При цьому кількість одиниць, відібраних з кожної групи, пропорційна кількості груп або середнім квадратичним відхиленням усередині типових груп.

Точність статистичного спостереження — ступінь відповідності величин якогось показника, отриманого за матеріалами статистичного спостереження, до дійсної його величини.

F-розподіл — це спільний закон розподілу двох взаємопов'язаних вибіркових дисперсій для випадкових величин x і y, кожна з яких розподілена нормально.

Факторна ознака — статистична ознака, яка впливає на іншу ознаку і зумовлює її зміну.

Факторне групування — групування, в якому групувальною ознакою є факторний показник, що впливає на зміну результативної ознаки. Якщо факторна ознака істотна, а кількість одиниць у групі досить велика, то інші умови в середньому по групі вирівнюватимуться і зміна результативного показника визначатиметься зміною факторного.

Факторний аналіз — перехід від вихідної інформації до узагальнених факторів, які є результатом їх первісної агрегації і лінійної комбінації.

Функціональний зв'язок — зв'язок, при якому кожному значенню факторної ознаки (аргументу), що характеризує певне явище, в усіх випадках відповідає одне або кілька значень результативної ознаки (функції).

Хемінгова відстань — використовується як міра відстані об'єктів, значення яких задано атрибутивними ознаками:

$$\rho(x_i, x_j) = \sum_{i,j}^{1,2,\ldots,k} \left| x_i^{1,2,\ldots,k} - x_j^{1,2,\ldots,k} \right|.$$

Центральна тенденція — це властивість значень досліджуваної ознаки групуватися навколо центра розподілу частот, статистичною характеристикою якого є середня величина.

Частість — це відносна величина структури, тобто питома вага частоти варіанта в загальній сумі частот. Також: відсотково виражена частота.

Часткові коефіцієнти кореляції — показники, які характеризують тісноту зв'язку результативної ознаки з однією факторною ознакою при умові, що інші факторні ознаки перебувають на постійному рівні. Парний коефіцієнт кореляції між результативною і факторною ознаками, як правило, не дорівнює відповідному частковому коефіцієнту.

Частота — кількість одиниць спостереження, що мають однакове значення ознаки. Іноді замість частот використовують частості.

ОСНОВНІ ФОРМУЛИ З МАТЕМАТИЧНОЇ СТАТИСТИКИ

Середня арифметична:

- проста $$\bar{x} = \frac{\sum x}{n} \quad (1)$$
- зважена $$\bar{x} = \frac{\sum xf}{\sum f} \quad (2)$$
- спосіб моментів $$\bar{x} = \frac{\sum\left(\frac{x-a}{h}\right)f}{\sum f}h + a \quad (3)$$

Середнє лінійне відхилення

$$d = \frac{\sum(x-\bar{x})}{n} \quad (4)$$

$$d = \frac{\sum(x-\bar{x})\,f}{\sum f} \quad (5)$$

Дисперсія:

- проста $$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n} \quad (6)$$
- зважена $$\sigma^2 = \frac{\sum(x-\bar{x})^2 f}{\sum f} \quad (7)$$
- спосіб моментів $$\sigma^2 = \frac{\sum\left(\frac{x-a}{h}\right)^2 f}{\sum f}\cdot h^2 - (\bar{x}-a)^2 \quad (8)$$
- спосіб моментів $$\sigma^2 = h^2(m_2 - m_1^2) \quad (9)$$

$$\sigma^2 = \overline{x^2} - \bar{x}^2 \quad (10)$$

Дисперсія альтернативної ознаки

$$\sigma^2 = pq \tag{11}$$

Математичні властивості дисперсії

$$1)\ \sigma^2_{x-a} = \sigma^2_x$$

$$2)\ \sigma^2_{\frac{x}{a}} = \sigma^2_x : a^2;\ \sigma_x = \sigma_{\frac{x}{a}} \cdot a \tag{12}$$

$$3)\ \sigma^2_a > \sigma^2_x$$

Середнє квадратичне відхилення:

$$\sigma = \sqrt{\sigma^2} \tag{13}$$

- просте

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \tag{14}$$

- зважене

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2 f}{\sum f}} \tag{15}$$

- альтернативної ознаки

$$\sigma = \sqrt{pq} \tag{16}$$

Показник варіації

$$V = \frac{\sigma}{\bar{x}} \cdot 100 \tag{17}$$

Момент першого порядку

$$m_1 = \frac{\sum \left(\frac{x - a}{h} \right) f}{\sum f} \tag{18}$$

Момент другого порядку

$$m_2 = \frac{\sum \left(\frac{x - a}{h} \right)^2 f}{\sum f} \tag{19}$$

Мода інтервального ряду розподілу

$$x_{мо} = x_0 + h \cdot \frac{f_{мо} - f_{мо-1}}{2f_{мо} - f_{мо+1} - f_{мо-1}} \tag{20}$$

$$x_{мо} = x_0 + h \cdot \frac{f_{мо} - f_{мо-1}}{(f_{мо} - f_{мо+1}) + (f_{мо} - f_{мо-1})} \tag{21}$$

Медіана інтервального ряду розподілу

$$x_{ме} = x_0 + h \frac{\frac{\sum f}{2} - S_{ме-1}}{f_{ме}} \tag{22}$$

Перший квартиль

$$Q_1 = x_{Q_1} + h \cdot \frac{\frac{\sum f}{4} - S_{Q_1-1}}{f_{Q_1}} \tag{23}$$

Другий квартиль

$$Q_2 = x_{ме} \tag{24}$$

Третій квартиль

$$Q_3 = x_{Q_3} + h \cdot \frac{\frac{\sum f}{4} - S_{Q_3-1}}{f_{Q3}} \tag{25}$$

Децилі

$$D = x_D + h \cdot \frac{\frac{1}{10}\sum f - S_{D-1}}{f_{D_n}} \tag{26}$$

Властивість мажорантності середніх

$$\bar{x}_{квадр} > \bar{x}_{арифм} > \bar{x}_{геом} > \bar{x}_{гарм} \tag{27}$$

Обсяг варіації

$$w = \sum (x - \bar{x})^2, \text{ або}$$
$$w = \sum (x - \bar{x})^2 f; \tag{28}$$

$$w = \sum x^2 - \frac{\left(\sum x\right)^2}{n}, \text{ або}$$

$$w = \sum x^2 - \frac{\left(\sum xf\right)^2}{\sum f};$$

$$w = \sum x^2 - n\overline{x^2}, \text{ або}$$

$$w = \sum x^2 f - f\overline{x^2}$$

Правило складання варіації

$$\sigma^2_{заг.} = \sigma^2_M + \sigma^2_B \quad (29)$$

Міжгрупова (систематична) варіація

$$w_M = \sum n_{гр}(\overline{x_{гр}} - \overline{x_з})^2$$

$$w_M = \frac{\sum\left(\sum x_{гр}\right)^2}{n} - \frac{\left(\sum x\right)^2}{N} \quad (30)$$

$$w_M = \frac{\sum\left(\sum x_{гр}\right)^2}{n} - N\overline{x_з^2}$$

Внутрішньогрупова (випадкова) варіація

$$w_B = \sum\sum(x - \overline{x_{гр}})^2 = \sum x^2 - \frac{\sum\left(\sum x_{гр}\right)^2}{n} \quad (31)$$

Загальний обсяг варіації

$$w_З = w_M + w_B \quad (32)$$

$$\sum(x - \overline{x_З})^2 = \sum n_{гр}(\overline{x_{гр}} - \overline{x_з})^2 + \sum\sum(x - \overline{x_{гр}})^2 \quad (33)$$

$$\sum x^2 - N\overline{x_З^2} = \frac{\sum\left(\sum x_{гр}\right)^2}{n} - N\overline{x_з^2} + \sum x^2 - \frac{\sum\left(\sum x_{гр}\right)^2}{n} \quad (34)$$

$$\sum x^2 - \frac{(\sum x^2)}{N} = \frac{\sum\left(\sum x_{гр}\right)^2}{n} - \frac{\left(\sum x\right)^2}{N} + \sum x^2 - \frac{\sum\left(\sum x_{гр}\right)^2}{n} \quad (35)$$

Помилка зміщення (поправочний коефіцієнт Бесселя)

$$\frac{n}{n-1} \quad (36)$$

Скоригована вибіркова дисперсія

$$S^2 = \sigma_0^2 \cdot \frac{n}{n-1} = \frac{\sum(x - \overline{x})^2}{n-1} \quad (37)$$

Середня помилка вибірки

$$m = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \quad (38)$$

Середня помилка вибірки для визначення частки

$$m = \sqrt{\frac{pq}{n}} \quad (39)$$

Правило трьох сигм

$$\frac{\varepsilon}{m} \le 3 \quad (40)$$

Нормоване відхилення

$$t = \frac{\varepsilon}{m} = \frac{\widetilde{x} - \overline{x}}{m} = \frac{x - \overline{x}}{\sigma} \quad (41)$$

Гранична помилка

$$\Delta = tm \quad (42)$$

Теорема Чебишова-Ляпунова:

- для середнього значення ознаки

$$p\left[(\widetilde{x} - \overline{x}) \le t\sqrt{\frac{\sigma^2}{n}}\right] = \Phi(t) \quad (43)$$

- для частки ознаки

$$p\left[(w-p) \le t\sqrt{\frac{pq}{n}}\right] = \Phi(t) \quad (44)$$

Інтервальна оцінка параметра

$$\Theta = \widetilde{\Theta} \pm \Delta \quad (45)$$

Надійний інтервал:

$$I_p = \left(\widetilde{\Theta} - \Delta; \widetilde{\Theta} + \Delta\right) \quad (46)$$

- для середнього значення ознаки

$$I_p = \left(\widetilde{x} - \Delta; \widetilde{x} + \Delta\right) \quad (47)$$

- для частості

$$I_p = \left(w - \Delta; w + \Delta\right) \quad (48)$$

Рівняння Гаусса-Лапласа

$$y = \frac{1}{\sigma_0\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} \quad (49)$$

якщо $\sigma_0 = 1$, то $f(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$

Оцінка результатів власне випадкової вибірки при визначенні середнього розміру ознаки:

- повторна

$$\Delta_x = t\sqrt{\frac{\sigma_0^2}{n}} \quad (50)$$

- безповторна

$$\Delta_x = t\sqrt{\frac{\sigma_0^2}{n} \cdot \left(1 - \frac{n}{N}\right)} \quad (51)$$

Оцінка результатів власне випадкової вибірки при визначенні частки ознаки:

- повторна

$$\Delta_p = t\sqrt{\frac{w-(1-w)}{n}} \quad (52)$$

- безповторна

$$\Delta_p = t\sqrt{\frac{w(1-w)}{n} \cdot \left(1 - \frac{n}{N}\right)} \quad (53)$$

Оцінка результатів серійної (гніздової) вибірки при визначенні середнього розміру ознаки:

- повторна

$$\Delta_x = t\sqrt{\frac{\sigma_c^2}{n_c}} \quad (54)$$

- безповторна

$$\Delta_x = t\sqrt{\frac{\sigma_c^2}{n_c} \cdot \left(1 - \frac{n_c}{N_c}\right)} \quad (55)$$

Оцінка результатів серійної (гніздової) вибірки при визначенні частки ознаки:

- повторна

$$\Delta_p = t\sqrt{\frac{w_c - (1-w_c)}{n_c}} \quad (56)$$

- безповторна

$$\Delta_p = t\sqrt{\frac{w_c(1-w_c)}{n_c} \cdot \left(1 - \frac{n_c}{N_c}\right)} \quad (57)$$

Оцінка результатів типової (районованої) вибірки при визначенні середнього розміру ознаки:

- повторна

$$\Delta_x = t\sqrt{\frac{\overline{\sigma_0^2}}{n}} \quad (58)$$

- безповторна

$$\Delta_x = t\sqrt{\frac{\overline{\sigma_0^2}}{n} \cdot \left(1 - \frac{n}{N}\right)} \quad (59)$$

Оцінка результатів типової (районованої) вибірки при визначенні частки ознаки:

- повторна
$$\Delta_p = t\sqrt{\frac{\overline{w-(1-w)}}{n}} \quad (60)$$
- безповторна
$$\Delta_p = t\sqrt{\frac{\overline{w-(1-w)}}{n}\cdot\left(1-\frac{n}{N}\right)} \quad (61)$$

Чисельність вибірки при визначенні середнього розміру ознаки:

- власне випадкова повторна вибірка
$$n = \frac{t^2\sigma_0^2}{\Delta_x^2} \quad (62)$$
- власне випадкова безповторна вибірка
$$n = \frac{t^2\sigma_0^2 N}{\Delta_x^2 N + t^2\sigma_0^2} \quad (63)$$
- серійна безповторна вибірка
$$n = \frac{t^2\sigma_c^2 N_c}{\Delta_x^2 N_c + t^2\sigma_c^2} \quad (64)$$
- типова безповторна вибірка
$$n = \frac{t^2\overline{\sigma_0^2}N}{\Delta_x^2 N + t^2\overline{\sigma_0^2}} \quad (65)$$

Чисельність вибірки при визначенні частки ознаки:

- власне випадкова повторна вибірка
$$n = \frac{t^2 w\cdot(1-w)}{\Delta_p^2} \quad (66)$$

- власне випадкова безповторна вибірка
$$n = \frac{t^2 w\cdot(1-w)N}{\Delta_p^2 N + t^2 w(1-w)} \quad (67)$$
- серійна безповторна вибірка
$$n = \frac{t^2 w_c\cdot(1-w_c)N_c}{\Delta_p^2 N_c + t^2 w_c(1-w_c)} \quad (68)$$
- типова безповторна вибірка
$$n = \frac{t^2\overline{w\cdot(1-w)}N}{\Delta_p^2 N + t^2\overline{w(1-w)}} \quad (69)$$

Дисперсія в малих вибірках ($n < 20$)
$$\sigma^2_{м.в.} = \frac{\sum(x-\widetilde{x})}{n-1} \quad (70)$$

Кількість ступенів свободи варіації
$$v = n-1 \quad (71)$$

Нульова гіпотеза
$$H_0 : x = a \quad (72)$$

Альтернативна гіпотеза
$$H_a : x \neq a \quad (73)$$

***F*-критерій Фішера-Снедекора**
$$F = S_1^2 : S_2^2 \quad (74)$$

Оцінка вірогідності двох середніх ($n_1 = n_2$):

- різниця між середніми двох вибірок
$$\Delta = \widetilde{x}_2 - \widetilde{x}_1 \quad (75)$$
- вибіркові дисперсії, скориговані на кількість ступенів свободи варіації
$$\sigma_1^2 = \frac{\sum x_1^2 - n\overline{x}_1^2}{n-1}, \quad \sigma_2^2 = \frac{\sum x_2^2 - n\overline{x}_2^2}{n-1} \quad (76)$$

- квадрати середніх помилок $$m_1^2 = \frac{\sigma_1^2}{n_1}, \quad m_2^2 = \frac{\sigma_2^2}{n_2} \qquad (77)$$

- середня помилка різниці двох середніх ($n_1 = n_2$) $$\overline{m}_{1-2} = \sqrt{m_1^2 + m_2^2} \qquad (78)$$

- фактичне значення t-критерію $$t_\phi = \frac{\widetilde{x}_2 - \widetilde{x}_1}{\overline{m}_{1-2}} \qquad (79)$$

- гранична помилка двох вибіркових середніх $$\Delta = t \cdot \overline{m}_{1-2} \qquad (80)$$

Оцінка вірогідності двох середніх ($n_1 \neq n_2$):

- суми квадратів відхилень двох вибірок $$W_1 = \sum x_1^2 - n_1 \widetilde{x}_1^2; \quad W_2 = \sum x_2^2 - n_2 \widetilde{x}_2^2 \qquad (81)$$

- об'єднана дисперсія двох вибірок $$\sigma_0^2 = \frac{\sum (x - \widetilde{x}_1)^2 + \sum (x - \widetilde{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{W_1 + W_2}{(n_1 - 1) + (n_2 - 1)} \qquad (82)$$

- середня помилка різниці двох середніх ($n_1 \neq n_2$) $$\overline{m}_{1-2} = \sqrt{\sigma_0^2 \frac{n_1 + n_2}{n_1 \cdot n_2}} \qquad (83)$$

- фактичне значення t-критерію $$t_\phi = \frac{\widetilde{x}_1 - \widetilde{x}_2}{\overline{m}_{1-2}} \qquad (84)$$

Оцінка вірогідності середньої різниці:

- середня різниця $$\overline{d} = \frac{\sum d}{n} = \frac{\sum (\widetilde{x}_2 - \widetilde{x}_1)}{n} \qquad (85)$$

- дисперсія окремих різниць d від середньої різниці $\overline{d}$ $$\sigma_d^2 = \frac{\sum d^2 - n\overline{d}^2}{n - 1} \qquad (86)$$

- середня помилка середньої різниці $$\overline{m}_d = \sqrt{\frac{\sigma_d^2}{n}} \qquad (87)$$

- фактичне значення t-критерію $$t_\phi = \frac{\overline{d}}{\overline{m}_d} \qquad (88)$$

- гранична помилка середньої різниці $$\Delta = t\overline{m}_d \qquad (89)$$

Перевірка відповідності емпіричного (фактичного) і теоретичного розподілу частот:

- нормоване відхилення $$t = \frac{x - \overline{x}}{\sigma} \qquad (90)$$

- теоретичні частоти $$f' = \Phi(t) \cdot \frac{N \cdot t}{\sigma} \qquad (91)$$

- фактичне значення критерію χ^2 $$\chi^2 = \sum \frac{(f - f')^2}{f'} \qquad (92)$$

- кількість ступенів свободи варіації $$v = (a - 1) \cdot (b - 1) \qquad (93)$$

Рівняння прямої лінії

$$y_x = a_0 + a_1 x \tag{94}$$

Система рівнянь для знаходження параметрів a_0 та a_1:

- 1 спосіб

$$\begin{aligned} \sum y &= na_0 + a_1 \sum x \\ \sum xy &= a_0 \sum x + a_1 \sum x^2 \end{aligned} \tag{95}$$

- 2 спосіб

$$a_0 = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \sum x \sum x}$$

$$a_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \sum x \sum x} \tag{96}$$

Коефіцієнт еластичності

$$E = a_1 \frac{\bar{x}}{\bar{y}} \tag{97}$$

Лінійний коефіцієнт парної кореляції

$$r_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2) \cdot (\overline{y^2} - \bar{y}^2)}} \tag{98}$$

Фактичне значення t-критерію Стьюдента

$$t_\phi = \frac{r}{m_r} \tag{99}$$

Середня помилка вибіркового коефіцієнта парної кореляції

$$m_r = \frac{1 - r^2}{\sqrt{n-2}} \tag{100}$$

Коефіцієнт детермінації

$$D = r^2 \tag{101}$$

Рівняння гіперболи

$$y_x = a_0 + a_1 \frac{1}{x} \tag{102}$$

$$\begin{aligned} \sum y &= na_0 + a_1 \sum \frac{1}{x} \\ \sum \frac{y}{x} &= a_0 \sum \frac{1}{x} + a_1 \sum \frac{1}{x^2} \end{aligned} \tag{103}$$

Індекс кореляції

$$i = \sqrt{\frac{\sigma^2_{y_x}}{\sigma^2_y}} = \sqrt{\frac{\sum (y_x - \bar{y})^2}{\sum (y - \bar{y})^2}} \tag{104}$$

Парабола другого порядку

$$y_x = a_0 + a_1 x + a_2 x^2 \tag{105}$$

$$\begin{aligned} \sum y &= na_0 + a_1 \sum x + a_2 \sum x^2 \\ \sum xy &= a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3 \\ \sum x^2 y &= a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4 \end{aligned} \tag{106}$$

$$\begin{aligned} \sum y &= na_0 + a_1 \sum (x - \bar{x}) + a_2 \sum (x - \bar{x})^2 \\ \sum y(x - \bar{x}) &= a_0 \sum (x - \bar{x}) + a_1 \sum (x - \bar{x})^2 + a_2 \sum (x - \bar{x})^3 \\ \sum (x - \bar{x})^2 y &= a_0 \sum (x - \bar{x})^2 + a_1 \sum (x - \bar{x})^3 + a_2 \sum (x - \bar{x})^4 \end{aligned} \tag{107}$$

$$\begin{aligned} \sum y &= na_0 + a_2 \sum (x - \bar{x})^2 \\ \sum y(x - \bar{x}) &= a_1 \sum (x - \bar{x})^2 \\ \sum (x - \bar{x})^2 y &= a_0 \sum (x - \bar{x})^2 + a_2 \sum (x - \bar{x})^4 \end{aligned} \tag{108}$$

Множинна регресія

$$y_x = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \tag{109}$$

Система рівнянь

$$\begin{aligned}
&\sum y = na_0 + a_1\sum x_1 + a_2\sum x_2 + \ldots + a_n\sum x_n \\
&\sum yx_1 = a_0\sum x_1 + a_1\sum x_1^2 + a_2\sum x_1x_2 + \ldots + a_n\sum x_1x_n \\
&\sum yx_2 = a_0\sum x_2 + a_1\sum x_1x_2 + a_2\sum x_2^2 + \ldots + a_n\sum x_2x_n^2 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&\sum yx_n = a_0\sum x_n + a_1\sum x_1x_n + a_2\sum x_2x_n + \ldots + a_n\sum x_n^2
\end{aligned} \qquad (110)$$

Парні коефіцієнти кореляцій

$$r_{yx_1} = \frac{\overline{x_1y} - \bar{x}_1\cdot\bar{y}}{\sqrt{\left(\overline{x_1^2} - \bar{x}_1^2\right)\cdot\left(\overline{y^2} - \bar{y}^2\right)}} \qquad (111)$$

$$r_{yx_2} = \frac{\overline{x_2y} - \bar{x}_2\cdot\bar{y}}{\sqrt{\left(\overline{x_2^2} - \bar{x}_2^2\right)\cdot\left(\overline{y^2} - \bar{y}^2\right)}} \qquad (112)$$

$$r_{x_1x_2} = \frac{\overline{x_1x_2} - \bar{x}_1\cdot\bar{x}_2}{\sqrt{\left(\overline{x_1^2} - \bar{x}_1^2\right)\cdot\left(\overline{x_2}^2 - \bar{x}_2^2\right)}} \qquad (113)$$

Коефіцієнт множинної детермінації

$$R^2 = \frac{\sigma^2_{обч}}{\sigma^2_{заг}} \qquad (114)$$

Дисперсія результативної ознаки, обчислена за рівнянням множинної регресії

$$\sigma^2_{обч} = \overline{y_x^2} - \bar{y}^2 \qquad (115)$$

Загальна дисперсія результативного показника

$$\sigma^2_{заг} = \overline{y^2} - \bar{y}^2 \qquad (116)$$

Коефіцієнт множинної кореляції

$$\sqrt{R^2} = \sqrt{\frac{\sigma^2_{обч}}{\sigma^2_{заг}}} \qquad (117)$$

Фактичне значення *t*-критерію

$$t_ф = \frac{R}{m_R} \qquad (118)$$

Середня помилка вибіркового коефіцієнта множинної кореляції

$$m_R = \frac{1-R^2}{\sqrt{n-m-1}} \qquad (119)$$

СПИСОК ЛІТЕРАТУРИ

1. Андерсен Т. Введение в многомерный статистический анализ. — М.: Физматгиз,1963. — 500 с.
2. Айвазян С. А, Енюков И. С, Мешалкин Л. Д. Прикладная статистика: Основы моделирования и первичной обработки данных. — М.: Финан сы и статистика, 1983. — 471 с.
3. Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ. — М.: Финансы и статистика, 1985. — 230 с.
4. Бернулли Я. О законе больших чисел. Под общей ред. Ю.В. Прохорова. — М.: Наука, 1986. — 176 с.
5. Бешелев С. Л., Гурвич Ф. Г. Математико-статистические методы экспертных оценок. — М.: Статистика, 1980. — 159 с.
6. Болч Б., Хуань К. Многомерные статистические методы для экономики. — М.: Статистика, 1979. — 316 с.
7. Боровиков В. STATISTICA: искусство анализа данных на компьютере. Для профессионалов. — СПб.: Питер, 2001. — 656 с.
8. Векслер Л. С. Статистический анализ на персональном компьютере // Мир ПК, № 2, 1992. — с. 89 – 97.
9. Головач А. В., Ерина А.М., Трофимов В.П. Критерии математической статистики в экономических исследованиях. — М.: Статистика, 1973.
10. Горкавий В.К. Статистика: Підручник. — К.: Вища шк., 1995. — 415 с.
11. Демиденко Е.З. Линейная и нелинейная регрессия. — М.: Финансы и статистика, 1981. — 302 с.
12. Джессен Р. Методы статистических обследований: Пер. с англ. / Под ред. и с предисл. Е.М. Четыркина. — М.: Финансы и статистика, 1985.
13. Дюк В. А., Мирошников А. И. Эволюция STATGRAPHICS // Мир ПК. — 1995. — № 12.
14. Кендэлл М. Ранговые корреляции. — М.: Статистика, 1975. — 212 с.

15. Кендэлл М., Стьюарт А. Статистические выводы и связи. — М.: Наука, 1973. — 899 с.
16. Крамер Г. Математические методы статистики. — М.: Мир, 1975. — 648 с.
17. Куценко С. П., Лазарєва С. Ф., Жаворонкова Г. В. Основні положення концепції та напрямки інформатизації суспільства // Машинна обробка інформації: Міжвідом. наук. Збірник. — КДЕУ, 1995. — Вип. 56. — 160 с.
18. Леман Э. Проверка статистическиъ гипотез. — М.: Наука, 1964. — 498 с.
19. Общая теория статистики: Статистическая методология в изучении комерческой деятельности: Учебник / А.И. Харламов, О. Э. Башина, В. Т. Бабурин и др.; Под ред. А. А. Спирина, О. Э. Башиной. — 4-е изд. — М.: Финансы и статистика, 1997. — 296 с.
20. Популярный экономико-статистический словарь-справочник / Под ред. И. И. Елисеевой. — М.: Финансы и статистика, 1993.
21. Про державну статистику: Закон України // Відомості Верховної Ради України, 1992, № 43.
22. Столяров Г. С., Ємшанов Д. Г., Ковтун Н. В. АРМ статистика: Навч. посібник. — К.: КНЕУ, 1999. — 268 с.
23. StatSoft, Inc. (1999). Электронный учебник по статистике. Москва, StatSoft. WEB: http://www.statsoft.ru/home/textbook/default.htm.
24. Теорія статистики: [Навч. посіб.] / Г.І. Мостовий, А. О. Дєгтяр, В. К. Горкавий, В. В. Ярова; За заг. ред. проф. Г. І. Мостового. — Х.: Вид-во ХарРІ УАДУ "Магістр", 2002. — 300 с.
25. Теория статистики: Учебник / Под ред. Р. А. Шмойловой. — 2-е изд., доп. и перераб.- М.: Финансы и статистика, 1998.
26. Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере / Под ред. В. Э. Фигурнова. — М.:ИНФРА — М, 1998. — 528 с.
27. Учебник StatSoft, Inc. (2001). Электронный учебник по промышленной статистике. Москва, StatSoft.WEB: http ://www .statsoft.ru / home / portal/ textbook_ind/default.htm.

28. Факторный, дискриминантный и кластерный анализ. — М.: Финансы и статистика, 1989. — 215 с.
29. Хартман Г. Современный факторный анализ. — М.: Статистика, 1972.
30. Холлендер М., Вульф Д. Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.
31. Чупров А. А. Основные проблемы теории корреляции. — М.: 1926. — с. 4 – 5.

ДОДАТКИ

Додаток 1 *Значення інтегралу імовірності при різних значеннях t*

t	Соті частки									
	0	1	2	3	4	5	6	7	8	9
0,0	0000	0080	0160	0239	0319	0399	0478	0558	0638	0717
0,1	0797	0876	0955	1034	1114	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2961	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4909	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6421	6476	6528	6579	6629	6679	6729	6778
1,0	6827	6875	6923	6970	7017	7063	7109	7154	7199	7423
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7995	8030
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8670	8715	8740	8764	8788	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9089
1,7	9108	9127	9146	9164	9182	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9425	9438	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9608	9615	9625	9634
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
3,0	99730	99739	99747	99755	99763	99771	99779	99786	99793	99800
3,1	99807	99813	99819	99825	99831	99837	99842	99847	99853	99858
3,2	99863	99867	99872	99876	99880	99884	99888	99892	99896	99900
3,3	99903									
3,6	99904									
3,9	99968									
4,4	9999892									
5,0	99999943									

Додаток 2 *Значення критерію Стьюдента при рівнях істотності 0,10, 0,05 і 0,01*

Ступені свободи варіації, ν	0,10	0,05	0,01	Ступені свободи варіації, ν	0,10	0,05	0,01
1	6,3138	12,706	63,657	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1825	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,6041	21	1,7207	2,0796	2,8314
5	1,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7530	2,1315	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2.8992				

Додаток 3 *Розподіл* χ^2

Ступені свободи варіації	Рівень істотності			Ступені свободи варіації	Рівень істотності		
	0,10	0,05	0,01		0,10	0,05	0,01
1	2,706	3,841	6,635	16	23,542	26,296	32,000
2	4,605	0,991	9,210	17	24,769	27,587	33,409
3	6,251	7,815	11,341	18	25,989	28,869	34,805
4	7,779	9,488	13,277	19	27,204	30,144	36,191
5	9,236	11,070	15,086	20	28,412	31,410	37,566
6	10,645	12,592	16,812	21	29,615	32,671	38,932
7	12,017	14,067	18,475	22	30,813	33,924	40,289
8	13,362	15,507	20,090	23	32,007	35,172	41,638
9	14,684	16,919	21,666	24	33,196	36,415	42,980
10	15,987	18,307	23,209	25	34,382	37,652	44,314
11	17,275	19,675	24,725	26	35,563	38,885	45,642
12	18,549	21,026	26,217	27	36,741	40,113	46,963
13	19,812	22,362	27,688	28	37,916	41,337	48,278
14	21,064	23,685	29,141	29	39,087	42,557	49,588
15	22,307	24,996	30,578	30	40,256	43,773	50,892

Додаток 4 *Значення імовірностей для критерію t-Стьюдента*

Кількість ступенів свободи варіації															
1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30
0,0	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000
0,1	063	071	073	075	076	077	077	077	078	078	078	079	080	080	080
0,2	126	140	146	149	151	152	153	154	154	155	155	156	156	156	158
0,3	186	208	216	221	224	226	227	228	229	230	231	232	232	233	236
0,4	242	272	284	290	294	297	299	300	302	303	304	305	306	306	311
0,5	295	333	347	357	362	365	368	369	371	373	375	376	377	377	383
0,6	344	391	409	419	425	430	433	435	437	439	441	443	444	444	452
0,7	389	444	466	477	485	490	493	496	498	502	504	505	507	508	516
0,8	430	492	518	531	540	546	550	553	556	558	562	564	565	566	576
0,9	467	537	566	581	591	597	602	606	608	613	616	618	619	621	632
1,0	500	577	609	626	637	644	649	653	657	661	664	667	669	670	683
1,1	530	614	648	667	679	687	692	697	700	705	709	711	713	715	729
1,2	558	647	684	704	716	725	731	736	739	745	748	751	753	755	770
1,3	583	677	716	737	750	759	765	770	774	780	784	788	789	791	806
1,4	605	704	744	766	780	789	796	801	805	811	815	818	821	822	838
1,5	626	728	769	792	806	716	823	828	832	838	842	846	848	850	866
1,6	644	749	792	815	830	839	846	852	856	862	866	870	872	874	890
1,7	661	769	812	836	850	860	867	872	877	884	887	890	893	895	911
1,8	677	786	830	854	868	878	885	890	895	901	905	908	910	912	982
1,9	692	802	846	870	884	894	901	906	910	916	920	923	925	927	943
2,0	705	816	861	884	898	908	914	919	923	929	933	936	938	940	954
2,1	717	829	873	896	910	920	926	931	935	940	944	947	949	951	964
2,2	728	841	885	907	921	930	936	941	945	950	954	956	958	960	972
2,3	739	852	895	917	930	935	945	950	953	958	961	964	966	967	979

Продовження додатку 4

Кількість ступенів свободи варіації															
1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30
2,4	749	862	904	926	938	947	953	957	960	965	968	970	972	973	984
2,5	758	870	912	933	946	953	959	963	966	970	973	975	977	978	988
2,6	766	878	920	940	952	959	965	968	971	975	978	980	981	982	991
2,7	774	886	926	946	957	964	969	973	976	979	982	984	985	986	993
2,8	782	893	932	951	962	969	973	977	979	983	985	987	988	989	995
2,9	789	899	937	956	966	973	977	980	982	986	988	989	990	991	996
3,0	795	905	942	960	970	976	980	983	985	988	990	991	992	993	997
3,1	901	910	947	962	973	979	983	985	987	990	992	993	994	994	998
3,2	807	915	951	967	976	981	985	987	989	992	992	994	995	995	999
3,3	813	919	954	970	979	984	987	989	991	993	994	995	996	996	999
3,4	818	923	958	973	981	986	989	991	992	994	995	996	997	997	999
3,5	823	927	961	975	983	987	990	992	993	995	996	997	997	998	1
3,6	828	931	963	977	984	989	991	993	994	996	997	997	998	998	
3,7	832	934	966	979	986	990	992	994	995	996	997	998	998	998	
3,8	836	937	968	981	987	991	993	995	996	997	998	998	999	999	
3,9	840	940	970	982	989	992	994	995	996	998	998	999	999	999	
4,0	844	943	972	984	990	993	995	996	997	998	998	999	999	999	
4,1	848	945	974	985	991	994	995	997	997	998	999	999	999	999	
4,2	851	948	975	986	992	994	996	997	998	999	999	999	999	1	
4,3	855	950	977	987	992	995	996	997	998	999	999	999	999		
4,4	858	952	978	988	993	995	997	998	998	999	999	999	1		
4,5	861	954	980	989	994	996	997	998	999	999	999	1			
4,6	864	956	981	990	994	996	998	998	999	999	1				
4,7	867	958	982	991	995	997	998	998	999	999					
4,8	969	959	983	991	995	997	998	999	999	999					
4,9	872	964	984	992	996	997	998	999	999	1					

Додаток 5 *Значення F при імовірності 0,95*

ν_2^*	ν_1 — ступінь свободи варіації для більшої дисперсії							
	1	2	3	4	5	6	7	8
1	161	200	216	225	230	234	237	239
2	18,51	19,0	19,2	19,3	19,3	19,3	19,4	19,4
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44
9	5,12	4,26	3,86	3,63	3,48	3,37	3,28	3,23
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70
15	4,54	3,68	3,30	3,09	2,90	2,79	2,71	2,64
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59
17	4,45	3,56	3,20	2,96	2,81	2,70	2,61	2,55
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36
25	4,24	3,38	2,99	2,76	2,60	2,49	2,40	2,34
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02
∞	3,84	2,99	2,60	2,37	2,21	2,10	2,01	1,94

Продовження додатку 5

ν_2^*	ν_1 — ступінь свободи варіації для більшої дисперсії						
	9	10	12	15	20	30	∞
1	241	242	244	246	248	250	254
2	19,4	19,4	19,4	19,4	19,5	19,5	19,5
3	8,81	8,79	8,74	8,70	8,66	8,62	8,53
4	6,00	5,94	5,91	5,86	5,80	5,75	5,63
5	4,77	4,74	4,68	4,62	4,56	4,50	4,36
6	4,10	4,06	4,00	3,94	3,87	3,81	3,67
7	3,68	3,64	3,57	3,51	3,44	3,38	3,23
8	3,39	3,35	3,28	3,22	3,15	3,08	2,93
9	3,18	3,14	3,07	3,01	2,94	2,86	2,71
10	3,02	2,98	2,91	2,85	2,77	2,70	2,54
11	2,90	2,85	2,79	2,72	2,65	2,57	2,40
12	2,80	2,75	2,69	2,62	2,54	2,47	2,30
13	2,71	2,67	2,60	2,53	2,46	2,38	2,21
14	2,65	2,60	2,53	2,46	2,39	2,31	2,13
15	2,59	2,54	2,48	2,40	2,33	2,25	2,07
16	2,54	2,49	2,42	2,35	2,28	2,19	2,01
17	2,49	2,45	2,38	2,31	2,23	2,15	1,96
18	2,46	2,41	2,34	2,27	2,19	2,11	1,92
19	2,42	2,38	2,31	2,23	2,16	2,07	1,88
20	2,39	2,35	2,28	2,20	2,12	2,04	1,84
21	2,37	2,32	2,25	2,18	2,10	2,01	1,81
22	2,34	2,30	2,23	2,15	2,07	1,98	1,78
23	2,32	2,27	2,20	2,13	2,05	1,96	1,76
24	2,30	2,25	2,18	2,11	2,08	1,94	1,73
25	2,28	2,24	2,16	2,09	2,01	1,92	1,71
26	2,27	2,22	2,15	2,06	1,99	1,90	1,69
27	2,25	2,20	2,13	2,06	1,97	1,88	1,67
28	2,24	2,19	2,12	2,04	1,96	1,87	1,65
29	2,22	2,18	2,10	2,03	1,94	1,85	1,64
30	2,21	2,12	2,09	2,01	1,93	1,84	1,62
40	2,12	2,08	2,00	1,92	1,84	1,74	1,51
60	2,04	1,99	1,92	1,84	1,75	1,65	1,39
120	1,96	1,91	1,83	1,75	1,66	1,55	1,25
∞	1,88	1,83	1,75	1,67	1,57	1,46	1,00

Примітка: ν_2^* — ступінь свободи варіації для меншої дисперсії

Додаток 6 *Критичні значення критерію Вілкоксона (сума рангів)*

Кількість пар	Рівень істотності		Кількість пар	Рівень істотності	
	0,05	0,01		0,05	0,01
7	2	0	14	21	13
8	4	0	15	25	16
9	6	2	16	30	20
10	8	3	17	35	23
11	11	5	18	40	28
12	14	7	19	46	32
13	17	10	20	52	38

Додаток 7 *Значення критерію z, що відповідають значенням коефіцієнта кореляції r від 0 до 0,99*

r	Значення z для сотих часток r									
	0	1	2	3	4	5	6	7	8	9
0.0	0.000	0.010	0.020	0.030	0.040	0.050	0.060	0.070	0.080	0.090
0.1	0.100	0.110	0.121	0.131	0.141	0.151	0.161	0.172	0.182	0.192
0.2	0.203	0.213	0.224	0.234	0.245	0.255	0.266	0.277	0.288	0.299
0.3	0.310	0.321	0.332	0.343	0.354	0.365	0.377	0.388	0.400	0.412
0.4	0.424	0.436	0.448	0.460	0.472	0.485	0.497	0.510	0.523	0.536
0.5	0.549	0.563	0.576	0.590	0.604	0.618	0.633	0.648	0.662	0.678
0.6	0.693	0.709	0.725	0.741	0.758	0.775	0.793	0.811	0.829	0.848
0.7	0.867	0.887	0.908	0.929	0.950	0.973	0.996	1.020	1.045	1.071
0.8	1.099	1.127	1.157	1.188	1.221	1.256	1.293	1.333	1.376	1.422
0.9	1.472	1.528	1.589	1.658	1.738	1.832	1.946	2.092	2.298	2.647
1.0	2.647	2.700	2.759	2.826	2.903	2.995	3.106	3.250	3.453	3.800

Додаток 8 *Критичні значення вибіркового коефіцієнта кореляції при різній кількості ступенів свободи варіації і рівнях значущості*

Кількість ступенів свободи	Рівень значущості		Кількість ступенів свободи	Рівень значущості	
	0,05	0,01		0,05	0,01
1	0,997	0,999	24	0,388	0,496
2	0,950	0,990	25	0,381	0,487
3	0,878	0,959	26	0,374	0,478
4	0,811	0,917	27	0,367	0,470
5	0,754	0,874	28	0,361	0,463
6	0,707	0,834	29	0,355	0,456
7	0,666	0,798	30	0,349	0,449
8	0,632	0,765	35	0,325	0,418
9	0,602	0,735	40	0,304	0,393
10	0,576	0,708	45	0,288	0,372
11	0,553	0,684	50	0,273	0,354
12	0,532	0,661	60	0,250	0,325
13	0,514	0,641	70	0,232	0,302
14	0,497	0,623	80	0,217	0,283
15	0,482	0,606	90	0,205	0,267
16	0,468	0,590	100	0,195	0,254
17	0,456	0,575	125	0,174	0,228
18	0,444	0,561	150	0,159	0,208
19	0,433	0,549	200	0,138	0,181
20	0,423	0,537	300	0,113	0,148
21	0,413	0,526	400	0,098	0,128
22	0,404	0,515	500	0,088	0,115
23	0,396	0,505	1000	0,062	0,081

Додаток 9 *Таблиця значення функції* $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}}$

t	0	1	2	3	4	5	6	7	8	9
0.0	3989	3989	3989	3989	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3725	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3588
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1392	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1013	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0438	0488	0478	0468	0459	0449
2,1	0444	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0203	0203	0138	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
4,0	0001	0001	0001	0000	0000	0000	0000	0000	0000	0000

Примітка: Всі значення збільшено в 10 000 разів

Видавничий дім "Професіонал"
пропонує:

Рекомендовано Міністерством освіти і науки України
для студентів вищих навчальних закладів

Курило И.А. и другие

МЕТОДЫ И АЛГОРИТМЫ РЕШЕНИЯ ЗАДАЧ РАСЧЕТА ЭЛЕКТРИЧЕСКИХ ЦЕПЕЙ В ПРИМЕРАХ И УПРАЖНЕНИЯХ

(На русском языке)

Навчальний посібник

"ВД "Професіонал",
2004. – 250 с.
Ціна: 17,00 грн.

На примерах решения конкретных задач представлены основные разделы современной теории электрических цепей, составляющие предмет теоретических основ электротехники. Объем теоретического материала, представленный в виде кратких физических схем и подробно изложенных алгоритмов, позволяет непосредственно перейти к решению как типовых, так и задач, выходящих за рамки основного курса указанной дисциплины. Оригинальный подбор задач и предложенные решения целиком способствуют этому. При этом, по форме изложения, реализуется естественный принцип выборочного прочтения и быстрого нахождения нужной информации.

Видавничий дім "Професіонал"
пропонує:

Моісєєв Л.М.
Ковальчук В.В.

МЕЗОМЕХАНІКА РЕГУЛЯРНОГО ДЕФОРМАЦІЙНОГО РЕЛЬЄФУ

Навчальний иосібник

"ВД "Професіонал",
2004. – 304 с.
Ціна: 31,00 грн. (з ПДВ)

У посібнику висвітлено оригінальні способи пластичного деформування кристалічних об'єктів, які дозволяють з достатньою повторюваністю результатів отримувати регулярні формозміни (дисипативні структури) на поверхні і усередині зразків правильної геомстричної форми. Наведено розроблений авторами науковий інструментарій для вивчення цього, по суті, нового фізичного явища у фізиці міцності і пластичності. Отримані експериментальні результати проаналізовано з позицій сінергетики деформованого твердого тіла. Створено новий науковий напрям, який дозволяє виявляти і вивчати ротаційні моди пластичності в кристалах.
Для наукових фахівців з фізики міцності й пластичності, а також для викладачів, аспірантів і студентів, котрі вивчають фізику твердого тіла.

Видавничий дім "Професіонал"
пропонує:

Рекомендовано Міністерством освіти і науки України
для студентів вищих закладів освіти

Савчук Л.О.
Гришпинюк О.В.

ІНФОРМАТИКА ТА КОМП'ЮТЕРНА ТЕХНІКА

Навчальний посібник

ВД "Професіонал",
2004. – 160 с.
Ціна: 15,00 грн.

У посібнику викладені стисли теоретичні відомості про операційні системи, про альтернативне вільно розповсюджуване програмне забезпечення офісної діяльності. Запропоновані завдання супроводжуються рисунками та ілюстраціями середовища операційної системи Linux та офісного пакету OpenOffice.org. Дотримані програмні вимоги, щодо викладення дисципліни "Інформатика та комп'ютерна техніка" в вищих навчальних закладах.

Видавничий дім "Професіонал"
пропонує:

Рекомендовано Міністерством освіти і науки України
для студентів вищих закладів освіти

Кириченко А.Ф.

ТЕОРЕТИЧНІ ОСНОВИ ІНЖЕНЕРНОЇ ГРАФІКИ

Підручник

ВД "Професіонал",
2004. – 496 с.
Ціна: 28,00 грн.

Підручник підготовлено відповідно до програми з інженерної графіки, затвердженої Міністерством освіти і науки України для студентів технічних спеціальностей. Видання подає не лише необхідні положення з теорії зображення та геометрії взагалі, воно розвиває у читача логічне та просторове мислення, його геометричний та графічний кругозір. Автор шукає шляхи кращого сприймання, тому кожна задача спочатку розглядається на просторовій моделі, а потім подається побудова комплексного креслення, наводить не тільки розв'язання задач, що випливають з теорії, але й аналізує результати, що здобуті різними методами.

Навчальне видання

Володимир Кузьмич Горкавий
Валентина Василівна Ярова

МАТЕМАТИЧНА СТАТИСТИКА

Керівник видавничих проектів О.С. Прокопчук
Дизайн обкладинки К. Р. Іщенко
Верстка С. О. Кіцно
Редактор Т. Є Кучеренко
Технічний редактор Ю.В. Романенко
Коректори М. А. Захарченко, Л. В. Шелудько

Формат 60x84/16. Підписано до друку 14.05.2004.
Друк офсетний. Папір офсетний. Гарнітура Таймс.
Наклад 1300 прим.

ТОВ "Видавничий дім "Професіонал"
Тел. (38-044) 268 93 95, 268 24 69, 269 86 73
Тел/факс (38-044) 269 63 30
e-mail: vdbook@ua.fm

Свідоцтво про внесення суб'єкту видавничої справи до Державного реєстру видавців, виготівників і розповсюджувачів видавничої продукції серія ДК №1533

Надруковано у видавництві "Фенікс".
Свідоцтво ДК №271 від 07.12.2000 р.
Зам. 24-238.

ГОРКАВИЙ Володимир Кузьмич - професор, завідувач кафедри статистики Харківського національного аграрного університету ім. В.В. Докучаєва, кандидат економічних наук. Автор першого в Україні підручника "Статистика" для аграрних ВНЗ (1995 р.), навчальних посібників "Статистика" (1994 р.), "Теорія статистики" (2002 р.). Переможець Всеукраїнського конкурсу навчальної літератури (1997 р.) та виставки-ярмарку науково-педагогічних ідей "Освіта Харківщини" (1998 р.). Автор 150 публікацій з проблем статистики та економічного аналізу.

ЯРОВА Валентина Василівна - доцент кафедри статистики Харківського національного аграрного університету ім. В.В. Докучаєва, кандидат економічних наук. Автор навчального посібника "Теорія статистики" (2002 р.).