**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ**

**ФАКУЛЬТЕТ ЗАХИСТУ РОСЛИН, БІОТЕХНОЛОГІЙ ТА ЕКОЛОГІЇ**

**Кафедра фізіології, біохімії рослин та біоенергетики**

**Нестерова Н.Г.**

# КУРС ЛЕКЦІЙ

з дисципліни

# «БІОМЕТРІЯ»

для студентів ОС «Бакалавр» очної форми навчання зі спеціальностей

**162 «Біотехнологія та біоінженерія»**

**229 «Громадське здоров'я»**

Київ 2024

УДК 57.087.1

Наведено лекційний курс з дисципліни «Біометрія». Вказано теоретичні основи та практичні методики, що дозволяють виявляти кількісні закономірності у біологічних явищах; принципи побудови математичних моделей біологічних явищ та процесів; навички та уміння комп'ютерної обробки експериментальних даних та правила коректного подання результатів досліджень для представлення у наукових публікаціях.

Для студентів аграрних та біологічних ВНЗ із напрямів підготовки «Біотехнологія» та «Харчові технології».

Рекомендовано до друку на засіданні кафедри фізіології, біохімії рослин та біоенергетики НУБіП України (протокол № 7 від 07 березня 2024 р.) та вченої ради факультету захисту рослин, біотехнологій та екології НУБіП України (протокол № 6 від 21 березня 2024 р.).

**Укладач:**
**Нестерова Наталія Георгіївна** – доц., к.с.-г.н., доцент кафедри фізіології, біохімії рослин та біоенергетики НУБіП України

**Рецензенти:**
**Лісовий Микола Михайлович** – проф., д.с.-г.н., професор кафедри екобіотехнології та біорізноманіття НУБіП України
**Лєндел Тарас Іванович** – доц., к.т.н., доцент кафедри автоматики та робототехнічних систем імені академіка І.І.Мартиненка НУБіП України

**НАВЧАЛЬНЕ ВИДАННЯ**

# БІОМЕТРІЯ

Конспект лекцій для студентів ОС «Бакалавр»
очної форми навчання зі спеціальностей
**162 «Біотехнологія та біоінженерія»**
**229 «Громадське здоров'я»**

Формат 60х90 1/16. Папір офсетний. Друк цифровий.
Наклад 100 прим. Ум. друк. арк. 4,5. Зам. № 692.
Друк ЦП «Компринт». Свідоцтво ДК №4131 від 04.08.2011 р.
м. Київ, вул. Предславинська, 28
528-05-42, 067-209-54-30
email: komprint@ukr.net

**Зміст**

***Вступ***

*«Статистика – це, безумовно, граматика наук,*
*бо науковець, перш за все, має прагнути до*
*самозниження у своїх судженнях і наводити аргументи,*
*які будуть абсолютно вірними як для кожного*
*окремого розуму, так і для його власного»*
*Карл Пірсон*

Якщо вихідні дані схильні до змін, то для отримання точнішого висновку абсолютно необхідно використовувати ідеї та техніки біометрії. Точність висновків, отриманих за допомогою біометричних процедур, зазвичай залежить від чотирьох найважливіших чинників: 1) уявлення та гнучкості мислення дослідника; 2) правильно підібраних методів отримання вибірки; 3) акуратної реєстрації вимірювань та врахування особливостей об’єктів вибірки; 4) коректного вибору та оформлення біометричних методів. Помилково думати, що біометричний метод може поліпшити некоректні дані – він буде тим ефективніше, чим точніше реєстратор дотримується перших трьох умов під час проведення експерименту.

Так, статистичний аналіз результатів біологічних досліджень дозволяє вирішувати декілька типів завдань:

- наочно представляти результати опису різнорідності об'єктів, що вивчаються;
- обґрунтовано (з певною ймовірністю помилки) приймати або не приймати припущення щодо наявності закономірностей, що відображаються у варіюванні вивченої величини;
- виявляти неявні закономірності, що приховані у варіюванні досліджуваних даних.

Проте, важливо уникати думки, що існує якась «особлива» біологічна статистика, яка принципово відрізняється від математичної в цілому. Безумовно, мінливість біологічних об'єктів має певні особливості, що відрізняють їх, наприклад, від мінливості фінансових показників або результатів технологічних процесів на виробництві. Це призводить до того, що набір методів, що використовуються в біології, відрізняється від аналогічних в інших сферах застосування статистики. Крім того, слід пам'ятати, що статистичне дослідження в біології не є самоціллю: воно підпорядковане завданням біологічного дослідження і не може бути повністю інтерпретовано поза біологічною проблемою, що вивчається. Проте, оскільки аналіз даних має бути підпорядкований логіці біологічного дослідження; то і сам експеримент саме має будуватися з урахуванням майбутнього аналізу. Збір емпіричних даних та постановка дослідів повинні заздалегідь враховувати, як саме буде організовано аналіз отриманих даних. Тобто, застосування статистики у біології неможливо повністю відмежувати від математичної як такої, бо воно все одно становить особливу галузь науки і особливий комплекс проблем з їх вирішенням. Для цієї галузі можна

використовувати термін, запропонований ще у 1899 р. Френсісом Гальтоном – біометрія. Водночас, враховуючи, що термін «біометрія» перехопили спеціалісти з ідентифікації особистості на підставі індивідуальних ознак, у багатьох випадках простіше використовувати і термін «біостатистика». На сьогодні ці два терміни позначають ідентичні показники і вільно співіснують паралельно один одному.

Об'єкти, які вивчає біологія, мають високий рівень унікальності. Практично у будь-якому біологічному феномені проявляються як загальні закономірності, так і вплив особливих обставин, що часто пов'язані з тієї чи іншою унікальністю біосистем. Це означає, що для біологічних досліджень дуже важливими є саме ті методи, які дозволяють побачити загальні закономірності, що проявляються мінливістю часткових проявів. Очевидно тому, біологи і зробили значніший внесок у розвиток статистики як самостійної науки. Результати робіт Френсіса Гальтона, Карла Пірсона, Рональда Фішера та багатьох інших складають важливу частину не лише біостатистики чи біометрії, а й математичної статистики в цілому.

В умовх сьогодення вже опрацьовано багато пакетів програм для математико-статистичної обробки даних. Першим і найпростішим пакетом математичного аналізу був і залишається *MICROSOFT EXCEL*, який дає змогу розуміти дані за допомогою запитів не статистичною мовою, ставити запитання безпосередньо про внесені дані, не набираючи складні формули. Крім того, такий пакет аналізу даних формує загальні візуальні зведення, тенденції та закономірності. Однак, така програма для сучасних дослідників вже дещо застаріла внаслідок своєї простоти, але якщо в арсеналі дослідних даних не об’ємний масив та запитом є прості математичні розрахунки – це стає перевагою. Особливого розповсюдження за останні десятиріччя набув потужний і зручний пакет програм системи *STATISTICA* для *Windows,* розроблений компанією StatSoft. Він має власну систему управління базами даних, сумісну з іншими системами, досить потужний пакет програм із графічним супроводом, графічний редактор тощо. Практичне освоєння цієї програми гарантує відсутність питань, пов’язаних з обробкою та форматуванням даних. Дещо складнішими є спеціалізовані програми для статистичного аналізу даних (LibreOffice Calc, RStudio, SPSS Statistics, EViews, AnalystSoft, StatPlus та інші), проте основними недоліками є висока вартість та досить складний алгоритм роботи, що передбачає високий теоретичний рівень дослідника. Даний лекційний посібник здебільшого орієнтований на використання статистичних програм *MICROSOFT EXCEL* та *STATISTICA* як бюджетних та функціональних надбудов, що вирішують більшість статистичних програм.

Лекційний курс розроблений з дисципліни «Біометрія» для студентів ОС «Бакалавр» спеціальностей 162 «Біотехнологія та біоінженерія» та 229 «Громадське здоров’я». Оскільки методи математико-статистичної обробки предметно не орієнтовані, то навчальне видання може використовуватися також студентами інших спеціальностей (біологів, лісівників, агробіологів, екологів та інших), аспірантів та викладачів суміжних галузей.

## *Лекція 1. Історія становлення як науки. Дані у біології*

**Біометрія** – розділ варіаційної статистики, за допомогою методів якої проводять обробку експериментальних даних і спостережень, а також планування кількісних експериментів у біологічних дослідженнях; а також власне наукова галузь, пов'язана з розробкою і використанням статистичних методів у наукових дослідженнях в медицині, біології та механіці тощо.

**Історія виникнення.** Біометрія як наука виникла у XIX столітті, головним чином, завдяки працям Френсіса Гальтона і Карла Пірсона. У 1920-30-х роках значний внесок у розвиток біометрії вніс Рональд Фішер. Однак, «винахідником» біометрії є Френсіс Гальтон (1822-1911 рр.). Спочатку Гальтон готувався стати лікарем, проте навчаючись у Кембриджському університеті, він захопився природознавством, метеорологією, антропологією, спадковістю і теорією еволюції. У його книзі, присвяченій природній спадковості, що було видана у 1889 р., вченим уперше було введено та використано слово *«biometry»*, а того ж року він розробив основи власне кореляційного аналізу. Так, Гальтон заклав основи нової науки і дав їй ім'я. Однак перетворив її у самостійну наукову дисципліну математик К. Пірсон (1857-1936). У 1884 р. Пірсон отримує посаду завідувача кафедри прикладної математики у Лондонському університеті, а в 1889 році знайомиться із Ф. Гальтоном та його роботами. Водночас, велику роль у житті Пірсона зіграв зоолог Уолтер Велдон. Допомагаючи йому в аналізі реальних зоологічних даних, у 1893 р. Пірсон ввів поняття середнього квадратичного відхилення і коефіцієнта варіації. Намагаючись математично оформити теорію спадковості Гальтона, Пірсон в 1898 р. розробляє основи множинної регресії. У 1903 р Пірсон розробив основи теорії пов'язаності ознак, а в 1905 р. опублікував основи нелінійної кореляції і регресії.

Наступний етап розвитку біометрії пов'язаний з ім'ям англійського статистика Рональда Фішера (1890-1962 рр.). Під час навчання у Кембриджському університеті Фішер знайомиться з працями Менделя і Пірсона. У 1913-1915 рр. Фішер працює статистом на одному з підприємств, а в 1915-1919 рр. – викладає фізику і математику в середній школі. З 1919 р. Фішер починає роботу статистом на дослідній сільськогосподарській станції у Ротамстеді, де він пропрацював до 1933 р. Потім з 1933 по 1943 р. Фішер працює професором у Лондонському університеті, а вже з 1943 по 1957 р. завідує кафедрою генетики у Кембриджі. Фішером були розроблені теорія вибіркових розподілів, методи дисперсійного та дискримінантного аналізу, теорії планування експериментів, метод максимальної правдоподібності та багато іншого, що становлять основу сучасної прикладної статистики та математичної генетики.

**Статистика** – галузь знань, у якій викладено загальні питання щодо збору, вимірювання та аналізу будь-яких масових статистичних (кількісних або якісних) даних. Слово *«статистика»* походить від латинського *status* – стан справ. У науку термін *«статистика»* ввів німецький учений Готфрід Ахенваль у 1746 році, запропонувавши замінити назву курсу

«Державознавство», що широко викладали у XVIII ст. в університетах Німеччини – на «Статистику», поклавши тим самим початок розвитку статистики і як науки, і як навчальної дисципліни. Водночас, слід зазначити, що статистичний облік вівся набагато раніше: проводився перепис населення у Стародавньому Китаї, здійснювалося порівняння військового потенціалу держав, вівся облік майна громадян в Стародавньому Римі тощо.

Статистика розробляє спеціальну методологію дослідження та обробки матеріалів: масові статистичні спостереження, метод угруповань, середніх величин, індексів, балансовий метод, метод графічних зображень і інші методи аналізу статистичних даних. Початок статистичної практики відноситься приблизно до часу виникнення держави. **Першою опублікованій статистичною інформацією** можна вважати глиняні таблички Шумерського царства (III - II тисячоліття до н. е.).

Спочатку під статистикою розуміли виключно опис економічного і політичного стану держави або його частини. Так, у 1792 р. було описано визначення: «статистика описує стан держави на даний час або у певний відомий момент у минулому». І на сьогодні діяльність державних статистичних служб цілком укладається у це визначення.

Однак поступово термін «статистика» став використовуватися значно ширше. За часів Наполеона Бонапарта, «статистика – це бюджет речей». Тим самим статистичні методи були визнані корисними не тільки для адміністративного управління, а й для застосування на рівні окремого підприємства. Згідно з формулюванням у 1833 р.: «мета статистики полягає у визначенні фактів у найстислішій формі». У 2-й половині XIX – початку XX століть сформувалася наукова дисципліна – **математичне статистика**, яка є частиною математики.

У XX столітті статистику часто розглядали, перш за все, як самостійну наукову дисципліну. Статистика являє собою сукупність методів і принципів, відповідно до яких проводиться збір, аналіз, порівняння, уявлення і інтерпретація числових даних. У 1954 р. математик, академік АН УРСР Борис Володимирович Гнеденко дав таке визначення: «Статистика складається з трьох розділів:

1. збір статистичних відомостей, тобто відомостей, що характеризують окремі одиниці будь-яких масових сукупностей;

2. статистичне дослідження отриманих даних, що полягає в з'ясуванні тих закономірностей, які можуть бути встановлені на основі даних масового спостереження;

3. розробка прийомів статистичного спостереження та аналізу статистичних даних. Останній розділ, власне, і становить зміст математичної статистики».

При цьому, термін «статистика» вживають ще в двох значеннях: по-перше, у побуті під «статистикою» часто розуміють набір кількісних даних про будь-яке явище або процес; по-друге, статистикою називають функцію від результатів спостережень, використовувану для оцінки характеристик і параметрів розподілів та перевірки гіпотез.

**Коротка історія статистичних методів.** Типові приклади раннього етапу застосування статистичних методів описані у Біблії. Там, зокрема, наводиться число воїнів у різних племенах. З математичної точки зору справа зводилася до підрахунку числа потраплянь значень спостережуваних ознак у певні градації.

Одразу після виникнення теорії ймовірностей (Паскаль, Ферма, XVII століття) імовірнісні моделі стали використовуватися при обробці статистичних даних. Наприклад, вивчалася частота народження хлопчиків і дівчаток, було встановлено відмінність ймовірності народження хлопчика від 0.5, аналізувалися причини того, що у паризьких притулках ця ймовірність не така, як у самому Парижі тощо.

У 1794 році (за іншими даними – у 1795 р) німецький математик Карл Гаусс формалізував один з методів сучасної математичної статистики – метод найменших квадратів. У XIX столітті помітний внесок у розвиток практичної статистики вніс бельгієць Кетле, який на основі аналізу великого числа реальних даних показав стійкість відносних статистичних показників, таких, як частка самогубств серед усіх смертей.

Перша третина XX століття пройшла під знаком **параметричної статистики**. Вивчалися методи, засновані на аналізі даних з параметричних сімей розподілів, описаних кривими сімейства Пірсона. Найпопулярнішим був нормальний розподіл. Для перевірки гіпотез використовувалися критерії Пірсона, Стьюдента та Фішера. Були запропоновані метод максимальної правдоподібності, дисперсійний аналіз, а також сформульовані основні ідеї планування експерименту.

Розроблену у першій третині XX століття теорію аналізу даних називають параметричною статистикою, оскільки її основний об'єкт вивчення – це вибірки з розподілів, що описуються лише одним або невеликим числом параметрів. Найзагальнішим є сімейство кривих Пірсона, що задаються чотирма параметрами. Як правило, не можна вказати будь-які вагомі причини, за якими розподіл результатів конкретних спостережень має входити в те чи інше параметричне сімейство. Винятки добре відомі: якщо ймовірна модель передбачає суму незалежних випадкових величин, то суму природно описувати нормальним розподілом; якщо ж в моделі розглядається як добуток таких величин, то підсумок, мабуть, наближається логарифмічно нормальним розподілом і так далі.

**Статистичні методи – методи аналізу статистичних даних.** Виділяють методи прикладної статистики, які можуть застосовувати в усіх областях наукових досліджень і будь-яких галузях народного господарства, та інші статистичні методи, застосування яких обмежене тією або іншою сферою. Маються на увазі такі методи, як статистичний приймальний контроль, статистичне регулювання технологічних процесів, надійність і випробування, планування експериментів тощо.

Статистичні методи аналізу даних застосовуються практично в усіх областях діяльності людини. Їх використовують завжди, коли необхідно

отримати і обґрунтувати будь-які судження про групу (об'єктів або суб'єктів) з деякою внутрішньою неоднорідністю.

Доцільно виділити **три види наукової і прикладної діяльності в області статистичних методів аналізу даних** (за ступенем специфічності методів, поєднаною з зануреною в конкретні проблеми):

а) розробка і дослідження методів загального призначення, без урахування специфіки галузі застосування;

б) розробка і дослідження статистичних моделей реальних явищ і процесів відповідно до потреб тієї чи іншої області діяльності;

в) застосування статистичних методів і моделей для статистичного аналізу конкретних даних.

**Прикладна статистика** – це наука про те, як обробляти дані довільної природи. Математичною основою прикладної статистики та статистичних методів аналізу є теорія ймовірностей і математична статистика.

Опис виду даних і механізму їх формування – початок будь-якого статистичного дослідження. Для опису даних застосовують як детерміновані, так і імовірнісні методи. За допомогою детермінованих методів можна проаналізувати лише ті дані, які є в розпорядженні дослідника. Наприклад, з їх допомогою отримані таблиці, розраховані органами офіційної державної статистики на основі представлених підприємствами і організаціями статистичних звітів. Перенести отримані результати на ширшу сукупність, використовувати їх для передбачення і управління можна лише на основі ймовірнісно-статистичного моделювання. Тому у суть математичної статистики часто включають лише методи, що спираються на теорію ймовірностей.

У простій ситуації **статистичні дані** – це значення деякої ознаки, властивої досліджуваним об'єктам. Значення можуть бути кількісними або являти собою вказівки на категорію, до якої можна віднести об'єкт. У другому випадку говорять про якісні ознаки.

При вимірюванні за кількома кількісними або якісними ознаками як статистичними даними про об'єкт отримуємо вектор. Його можна розглядати і як новий вид даних. В такому випадку вибірка складається з набору векторів. Є частина координат – числа, а частина – якісні (категоризовані) дані, або вектор різнотипних даних.

Одним з елементів вибірки, тобто одним виміром, може бути і функція в цілому. Наприклад така, що описує динаміку показника, тобто його зміни у часі, – електрокардіограма хворого або амплітуда биття вала двигуна. Також існує тимчасовий ряд, що описує динаміку показників певної організації. Тоді вибірка складається із набору функцій.

Елементами вибірки можуть бути і інші математичні об'єкти. Наприклад, бінарні відносини. Так, під час опитувань експертів часто використовують впорядкування (ранжування) об'єктів експертизи – зразків продукції, інвестиційних проектів, варіантів управлінських рішень. Залежно від регламенту експертного дослідження елементами вибірки можуть бути

різні види бінарних відносин (упорядкування, розбиття, толерантності), безлічі, нечіткі множини тощо.

Отже, математична природа елементів вибірки в різних завданнях прикладної статистики може бути найрізноманітнішою. Однак можна виділити два класи статистичних даних – **числові і не числові**. Так, прикладна статистика розбивається на дві частини – числову статистику і не числову статистику.

Числові статистичні дані – це числа, вектори, функції. Їх можна складати або множити на коефіцієнти. Тому у числовій статистиці велике значення мають різноманітні суми. **Математичний апарат аналізу сум випадкових елементів вибірки** – це (класичні) закони великих чисел і центральні граничні теореми.

Не числові статистичні дані – це категоризовані дані, вектори різнотипних ознак, бінарні відносини, безлічі, нечіткі множини та ін. Їх не можна складати і множити на коефіцієнти, тому не має сенсу говорити про суми не числових статистичних даних. Вони є елементами виключно не числових математичних просторів (множин). Математичний апарат аналізу не числових статистичних даних заснований на використанні відстаней між елементами (а також заходів близькості, показників відмінності) в таких просторах. За допомогою відстаней визначаються емпіричні та теоретичні середні, доводяться закони великих чисел, будуються непараметричні оцінки щільності розподілу ймовірностей, вирішуються завдання діагностики і кластерного аналізу тощо.

У прикладних дослідженнях використовують статистичні дані різних видів. Це пов'язано, зокрема, зі способами їх отримання. Наприклад, якщо випробування деяких технічних пристроїв тривають до певного моменту часу, то отримуємо так звані **цензуровані дані**, що складаються з набору чисел – тривалості роботи ряду пристроїв до відмови, і інформації про те, що інші пристрої продовжували працювати в момент закінчення випробування. Цензуровані дані часто використовуються при оцінці і контролі надійності технічних пристроїв.

Теорія статистичних методів націлена на вирішення реальних завдань. Тому в ній постійно виникають нові постановки математичних завдань аналізу статистичних даних, розвиваються і обґрунтовуються нові методи. Обґрунтування часто проводиться математичними засобами, тобто шляхом доведення теорем. Велику роль відіграє методологічна складова – як саме ставити завдання, які припущення прийняти з метою подальшого математичного вивчення. Суттєвою є роль сучасних інформаційних технологій, зокрема, комп'ютерного експерименту.

Розвиток обчислювальної техніки у другій половині XX століття справила значний вплив на статистику. Раніше статистичні моделі були представлені переважно лінійними моделями. Збільшення швидкодії ЕОМ і розробка відповідних чисельних алгоритмів послужило причиною підвищеного інтересу до таких нелінійних моделей, як штучні нейронні

мережі, і привело до розробки складних статистичних моделей, наприклад узагальнена лінійна модель і ієрархічна модель.

**Статистичне спостереження** – це масове (охоплює велику кількість випадків прояву досліджуваного явища для отримання правдивих статистичних даних) планомірне (проводиться за розробленим планом, що включає питання методології, організації збору і контролю за достовірністю інформації), систематичне (проводиться систематично, або безперервно, або регулярно) , науково організоване (для підвищення достовірності даних, яка залежить від програми спостереження, змісту анкет, якості підготовки інструкцій) спостереження за явищами і процесами соціально-економічного життя, яке полягає в зборі та реєстрації окремих ознак у кожної одиниці сукупності.

**Етапи статистичного спостереження:**

1. Підготовка до статистичного спостереження (рішення науково-методичних і організаційно-технічних питань).
   - визначення мети і об'єкта спостереження;
   - визначення складу ознак які підлягають вивченню;
   - розробка документів для збору даних;
2. Збір інформації.
   - безпосереднє заповнення статистичних формулярів (бланки, анкети);
   - застосування стандартних методів збору (пробні майданчики, пастки Геро тощо).

**Статистична інформація** – це первинні дані про предмет вивчення, що формуються в процесі статистичного спостереження, які потім піддаються систематизації, зведенню, аналізу та узагальнення.

3. Первинна обробка даних.
4. Статистичний аналіз обробленої інформації.
5. Розробка пропозицій і рекомендацій щодо вдосконалення статистичного спостереження полягає в аналізі причин, які призвели до неправильного заповнення статистичних формулярів та розробці відповідних пропозицій щодо вдосконалення спостереження.

У результаті статистичного спостереження повинна бути отримана об'єктивна, порівняльна, повна інформація, що дозволяє на подальших етапах дослідження забезпечити науково-обґрунтовані висновки про характер і закономірності розвитку досліджуваного явища.

**Види статистичного спостереження.** Статистичні спостереження поділяються на види за такими ознаками:

- за часом реєстрації даних;
- за повнотою охоплення одиниць сукупності;

**Види статистичного спостереження за часом реєстрації. Поточне (безперервне) спостереження** проводиться для вивчення поточних явищ і процесів. Реєстрація фактів здійснюється згідно з термінами їх завершення.

**Перерване спостереження** проводиться за необхідності, при цьому допускаються тимчасові розриви у реєстрації даних.

- Періодичне спостереження проводиться через порівняно рівні інтервали часу.
- Одночасне спостереження здійснюється без дотримання суворої періодичності його проведення.

По повноті охоплення одиниць сукупності розрізняють наступні види статистичного спостереження:

**Суцільне спостереження** – являє собою збір і отримання інформації про всі одиниці досліджуваної сукупності.

**Не суцільне спостереження** – засноване на принципі випадкового відбору одиниць досліджуваної сукупності, при цьому у вибірковій сукупності повинні бути представлені усі типи одиниць, наявних у сукупності.

Не суцільне спостереження підрозділяється на:

- Вибіркове спостереження – засновано на випадковому відборі одиниць, які піддаються спостереженню.
- Монографічне спостереження – полягає в обстеженні окремих одиниць сукупності, що характеризуються рідкісними якісними властивостями.
- Метод основного масиву – полягає у вивченні найістотніших, найбільших одиниць сукупності, що мають за основною ознакою найбільшу питому вагу у досліджуваній сукупності.
- Метод моментних спостережень – полягає у проведенні спостережень через випадкові або постійні інтервали часу з відмітками про стан досліджуваного об'єкта в той чи інший момент часу.

**Способи статистичного спостереження. Безпосереднє статистичне спостереження** – спостереження, при якому самі реєстратори шляхом безпосереднього виміру, зважування, підрахунку встановлюють факт, що підлягає реєстрації. **Документальне спостереження** засновано на використанні різного роду документів облікового характеру. **Опитування** полягає в отриманні необхідної інформації безпосередньо від респондента.

На сьогодні сформульовано наступні види опитування:

- **Експедиційне** – реєстратори отримують необхідну інформацію від опитуваних осіб і самі фіксують її в формулярах.
- **Самореєстрація** – формуляри заповнюються самими респондентами, реєстратори тільки роздають бланки і пояснюють правила їх заповнення.
- **Кореспондентське** – відомості до відповідних органів повідомляє штат добровільних кореспондентів.
- **Анкетне** – збір інформації здійснюється у вигляді анкет, які представляють собою спеціальні запитальники, спосіб зручний у випадках, коли не потрібна висока точність результатів.

- **Явочне**– полягає в наданні відомостей до відповідних органів в явочному порядку.

Залежно від причин виникнення розрізняють помилки реєстрації та помилки репрезентативності. Помилки реєстрації характерні як для суцільного, так і для не суцільного спостереження, а помилки репрезентативності – тільки для не суцільного спостереження. Помилки реєстрації, як і помилки репрезентативності, можуть бути випадковими і систематичними.

**Помилки реєстрації** – є відхиленням між значенням показника, отриманого в ході статистичного спостереження, і його фактичним значенням. Помилки реєстрації бувають випадковими (результат дії випадкових чинників – переплутані рядки наприклад) і систематичними (проявляються постійно).

**Помилки репрезентативності** – виникають, коли відібрана сукупність недостатньо точно відтворює вихідну сукупність. Характерні для не суцільного спостереження і полягають у відхиленні величини показника досліджуваної частини сукупності від його величини у генеральній сукупності.

**Випадкові помилки** – є результатом дії випадкових чинників.

**Систематичні помилки** – завжди мають однакову спрямованість до збільшення або зменшення показника по кожній одиниці спостереження, внаслідок чого значення показника за сукупністю в цілому буде включати накопичену помилку.

**Способи контролю:**

- Лічильний (арифметичний) – перевірка правильності арифметичного розрахунку.
- Логічний – заснований на смисловому взаємозв'язку між ознаками.

**Статистична сукупність** – безліч одиниць, що володіють масовістю, типовістю, якісною однорідністю і наявністю варіації. Статистична сукупність складається з матеріально існуючих об'єктів і є об'єктом статистичного дослідження. **Одиниця сукупності** – кожна конкретна одиниця статистичної сукупності. Одна і та ж статистична сукупність може бути однорідною за однією ознакою і неоднорідною за іншою.

**Якісна однорідність** – схожість всіх одиниць сукупності за будь-якою ознакою і несхожість за всіма іншими.

У статистичній сукупності відмінності однієї одиниці сукупності від іншої частіше мають кількісну природу. Кількісні зміни значень ознаки різних одиниць сукупності або різноманіття і мінливість величини ознак в окремих одиниць сукупності називається **варіацією**. **Варіація ознаки** – кількісна зміна ознаки (для кількісної ознаки) при переході від однієї одиниці сукупності до іншої.

**Ознака** – це властивість, характерна риса або інша особливість одиниць, об'єктів і явищ, яка може спостерігатися або вимірюватися. Ознаки поділяються на кількісні і якісні. **Атрибутивні (якісні) ознаки** не піддаються

числовому вираженню (склад населення за родом). **Кількісні ознаки** мають числовий вираз (склад населення за віком).

**Показник** – це узагальнююча кількісно-якісна характеристика будь-якої властивості одиниць або сукупності в цілому в конкретних умовах часу і місця.

**Система показників** – це сукупність показників, що всебічно відображають досліджуване явище. Наприклад, вивчається приріст корів:

- Ознака – вага.
- Статистична сукупність – корови ферми.
- Одиниця сукупності – кожна корова.
- Якісна однорідність – корови одного віку.
- Варіація ознаки – ряд цифр.

**Генеральна сукупність і вибірка з неї.** Основу статистичного дослідження становить безліч даних, отриманих в результаті вимірювання одного або декількох ознак. Реально відтворювана сукупність об'єктів, що статистично представлена рядом спостережень $^{x1,\ x2...xn}$ випадкової величини X, є **вибіркою**, а гіпотетично існуюча (домислювана) – **генеральною сукупністю**. Генеральна сукупність може бути кінцевою (число спостережень **N = const**) або нескінченної (**N = ∞**), а вибірка з генеральної сукупності – це завжди результат обмеженого ряду спостережень. Число спостережень, що утворюють вибірку, називається **обсягом вибірки**. Якщо обсяг вибірки досить великий (**n → ∞**) вибірка вважається **великою**, в іншому випадку вона називається **вибіркою обмеженого обсягу**. Вибірка вважається **малою**, якщо при вимірюванні одновимірної випадкової величини обсяг вибірки не перевищує 30 (**n <= 30**), а при вимірюванні одночасно декількох (k) ознак в багатовимірному просторі відношення n до k не перевищує 10 (**n / k <10**). Вибірка утворює **варіаційний ряд**, якщо її члени є **порядковими статистиками**, тобто вибіркові значення випадкової величини X впорядковані за зростанням (ранжовані), значення ж ознаки називаються **варіантами**.

**Основні способи організації вибірки.** Достовірність статистичних висновків і змістовна інтерпретація результатів залежить від **репрезентативності вибірки**, тобто повноти та адекватності уявлення властивостей генеральної сукупності по відношенню до якої цю вибірку можна вважати представницькою. Вивчення статистичних властивостей сукупності можна організувати двома способами: за допомогою суцільного і не суцільного спостереження.

**Суцільне спостереження** передбачає обстеження всіх одиниць досліджуваної сукупності, а **не суцільне (вибіркове) спостереження** – тільки його частини.

Існує **п'ять основних способів організації вибіркового спостереження:**

1. простий випадковий відбір, при якому *n-об'єктів* випадково витягуються з генеральної сукупності *N-об'єктів* (наприклад за допомогою таблиці або

датчика випадкових чисел), причому кожна з можливих вибірок мають рівну ймовірність. Такі вибірки називаються **власне-випадковими**;

2. простий відбір за допомогою регулярної процедури здійснюється механічною складовою (наприклад, дати, дня тижня, номера квартири, літери алфавіту та ін.), а отримані таким способом вибірки називаються **механічними**;
3. стратифікований відбір полягає в тому, що генеральна сукупність обсягу *N* підрозділяється на підсукупності або шари (страти) обсягу *N1,N2…Nr* так що *N1+N2+…+Nr=N*. Страти є однорідними об'єктами з точки зору статистичних характеристик (наприклад, населення ділиться на страти за віковими групами або соціальним статусом; підприємства – по галузях тощо). В цьому випадку вибірки називаються **стратифікованими** (іншими словами, розшарованими, типовими, районованими);
4. методи серійного відбору використовуються для формування серійних або гніздових вибірок. Вони зручні в тому випадку, коли необхідно обстежити відразу "блок" або серію об'єктів (наприклад, партію товару, продукцію певної серії або населення під час територіально-адміністративному поділі країни). Відбір серій можна здійснити власне-випадковим або механічним способом. При цьому проводиться суцільне обстеження певної партії товару або цілої територіальної одиниці (житлового будинку або кварталу);
5. комбінований (ступінчастий) відбір може поєднувати в собі відразу кілька способів відбору (наприклад, стратифікований і випадковий або випадковий і механічний); така вибірка називається **комбінованою**.

**Види відбору.** За видами розрізняються **індивідуальний, груповий і комбінований відбір. При індивідуальному відборі** до вибіркової сукупності відбираються окремі одиниці генеральної сукупності, **при груповому відборі** – якісно однорідні групи (серії) одиниць, а **комбінований відбір** передбачає поєднання першого і другого видів.

За методом відбору розрізняють **повторювану і неповторювану вибірку**. **Неповторюваним** називається відбір, при якому одиниця, що потрапила у вибірку не повертається у вихідну сукупність і надалі у відборі участі не бере; при цьому чисельність одиниць генеральної сукупності *N* скорочується в процесі відбору. **При повторюваному відборі** одиниця , що потрапила у вибірку після реєстрації повертається в генеральну сукупність і таким чином зберігає рівну можливість поряд з іншими одиницями бути використаною у подальшій процедурі відбору; при цьому чисельність одиниць генеральної сукупності *N* залишається незмінною (такий метод в соціально-економічних дослідженнях застосовується рідко). Однак, при великому *N* ($N \to \infty$) формули для неповторного відбору наближаються до аналогічних для повторного і на практиці частіше використовуються останні ($N = const$).

За своєю природою розподіли бувають **безперервними і дискретними**. Найвідомішим безперервним розподілом є **нормальний**. Залежно від виду розподілу і від способу відбору одиниць сукупності по-різному

обчислюються характеристики параметрів розподілу: **теоретичний та емпіричний розподіл.**

**Долею вибірки *Кп*** називається відношення числа одиниць вибіркової сукупності до числа одиниць генеральної сукупності:

$$kn = n / N.$$

**Вибіркова частка *w*** – це відношення одиниць, які мають досліджувану ознаку *X* до обсягу вибірки n:

$$w = Nn / n.$$

*Приклад.* У партії товару, що містить 1000 од., при 5% вибірці доля вибірки *Кп* у абсолютній величині становить 50 од. (N = N * 0,05); якщо ж в цій вибірці виявлено 2 бракованих вироби, то вибіркова частка браку *w* складе 0,04 (w = 2/50 = 0,04 або 4%).

Так як вибіркова сукупність відмінна від генеральної, то виникають **помилки вибірки.**

**Шкали вимірювань. Стан об'єкта** зазвичай оцінюється за критеріями. У якості критеріїв можуть виступати: виживання тварин, ступінь інтоксикації, збереження життєво важливих функцій і т.д. **Оцінки** вимірюються в тій чи іншій шкалі. **Шкала** (*умовно кажучи, шкала – це безліч можливих значень оцінок за критеріями*) – числова система, в якій відносини між різними властивостями досліджуваних явищ чи процесів переведені у властивості такої чи інакшої множини, як правило – множини чисел.

Розрізняють декілька типів **шкал**:

По-перше, можна виділити **дискретні шкали** (в яких безліч можливих значень оцінюваної величини є показником скінченним – наприклад, оцінка у балах – «1», «2», «3», «4», «5») і **безперервні шкали** (наприклад, концентрація речовини в моль / л або активність ферменту у сироватці крові в мКАТ / л).

По-друге, виділяють **шкали відносин, інтервальні шкали, порядкові (рангові) шкали і номінальні шкали (шкали найменувань).**

**Шкала відносин** – найпотужніша шкала. Вона дозволяє оцінювати, у скільки разів один вимірюваний об'єкт більший (менший) за інший, що є еталоном. Для шкал відносин існує стандартний початок відліку (нуль), але немає стандартної одиниці вимірювань. Шкалами відносин вимірюються майже всі фізичні величини – час, лінійні розміри, площі, обсяги, сила струму, потужність і т.д. У медичних та біологічних дослідженнях шкала відносин матиме місце, наприклад, коли вимірюється час появи тієї чи іншої ознаки після певного впливу (поріг часу, в секундах, хвилинах), інтенсивність впливу до появи якої-небудь ознаки (поріг сили впливу в вольтах, рентгенах і т.п.). Природно, до шкали відносин приймають усі дані у біохімічних і електрофізіологічних дослідженнях (концентрації речовин, вольтаж, тимчасові показники електрокардіограми тощо). Сюди ж, наприклад, відносяться і кількість правильно чи неправильно виконаних «завдань» в різних тестах з вивчення вищої нервової діяльності у тварин.

**Шкала інтервалів** застосовується досить рідко і характеризується тим, що для неї не існує ані стандартного початку відліку, ані стандартної одиниці

виміру. Прикладом шкали інтервалів є шкала температур за Цельсієм, Реомюр або Фаренгейтом. Шкала Цельсія, як відомо, була встановлена наступним чином: за нуль була прийнята точка замерзання води, за 100 градусів – точка її кипіння, і, відповідно, інтервал температур між замерзанням і кипінням води поділений на 100 рівних частин. Тут вже твердження, що температура $300^0$С в три рази більше, ніж $100^0$С, буде невірним. У шкалі інтервалів зберігається відношення довжин інтервалів. Можна сказати так: температура в $300^0$С відрізняється від температури в $200^0$С в два рази сильніше, ніж температура в $150^0$С відрізняється від температури в $100^0$С.

**Порядкова шкала (шкала рангів)** – шкала, щодо значень якої вже не можна говорити ні про те у скільки разів вимірювана величина більше (менше) іншої, ні на скільки вона більша (менша). Така шкала тільки впорядковує об'єкти, приписуючи їм ті чи інші бали (результатом вимірювань є не «суворе» впорядкування об'єктів). Наприклад, так побудована шкала твердості мінералів Мооса: взятий набір 10 еталонних мінералів для визначення відносної твердості методом дряпання. За 1 прийнято тальк, за 2 – гіпс, за 3 – кальцит і так далі до 10 – алмаз. Будь-якому мінералу однозначно може бути приписана відповідна певна твердість. Якщо досліджуваний мінерал, допустимо, дряпає кварц (7), але не дряпає топаз (8), то відповідно його твердість буде дорівнює 7. Аналогічно побудовані шкали сили вітру Бофорта і землетрусів Ріхтера. Шкали порядку широко використовуються в педагогіці, психології, медицині та інших науках, не настільки точних, як, скажімо, фізика та хімія. Зокрема, повсюдно поширена шкала шкільних оцінок в балах (п'ятибальна, дванадцятибальна і т.д.) може бути віднесена до шкали порядку. У медико-біологічних дослідженнях шкали порядку зустрічаються часто і майстерно замасковані. Наприклад, для аналізу згортання крові використовується Тромботест: 0 – відсутність згортання протягом часу тесту (а через хвилину після?), 1 – «слабкі нитки», 2 – желеподібний згусток, 3 – згусток, що легко деформується, 4 – щільний, пружний, 5 – щільний, що займає весь об'єм і т.п. Зрозуміло, що інтервали між цими погано відзначеними окремо і дуже суб'єктивними позиціями є довільними. У цьому випадку фраза «Тромботест у досліджуваних тварин підвищувався у середньому з 3,3 до 3,7» виглядає абсурдною. Безліч подібних шкал все ще зустрічається в експериментальній токсикології, експериментальної хірургії та експериментальній морфології. Окремим випадком порядкової шкали є **дихотомічна шкала**, в якій є лише дві впорядковані градації – наприклад, «вижив після експерименту» або «не вижив».

**Шкала найменувань (номінальна шкала)** фактично вже не пов'язана з поняттям «величина» і використовується тільки з метою відрізнити один об'єкт від іншого: номер тварини в групі або присвоєний йому унікальний шифр тощо.

## *Лекція 2. Елементи теорії планування досліджень*

**Цілі і завдання науки. Предмет біометрії** – вивчення властивостей масових явищ у біології. Ці явища зазвичай уявляються складними внаслідок різноманітності (варіювання) окремих індивідуумів або одиниць. Щоб отримати правильне уявлення про досліджувані властивості масових явищ і дати їм певні кількісні оцінки, їх піддають спільному розгляду та аналізу. Окремі одиниці або індивідууми, що володіють деякими загальними властивостями, об'єднують в **сукупності**. Спостережувані одиниці називають **варіантами (даними, датами)**, а утворену сукупність одиниць – **статистичною сукупністю**. Статистична сукупність може бути утворена за однією або декількох ознаках. Вона може складатися з однієї або декількох груп, що є однорідними щодо досліджуваної властивості. Однак часто буває доцільним поділити окремі спостережувані одиниці на групи для досягнення більшої однорідності їх усередині цих груп.

Теорію і методи вивчення властивостей масових явищ, обчислення та аналізу їх кількісних характеристик вивчає **біометрія**. Метод вивчення масових явищ заснований на теорії ймовірностей. Теорія ймовірностей встановлює закономірності подій, що виникають випадково і мають назву **випадкових**. Статистика передбачає аналіз масових явищ, що мають також випадковий характер в розподілі значень окремих одиниць, що складають явище. **Центральним завданням біометрії** як методу дослідження є висновки, що виходять за рамки вивченого матеріалу, тобто висновки про властивості статистичних сукупностей, беручи до уваги і невивчену їх частину. Всю статистичну сукупність, щодо якої роблять статистичні узагальнення та висновки, називають **загальною**, або **генеральною сукупністю**, а її частину, охоплену безпосереднім спостереженням, називають **вибірковою сукупністю.** Варіаційна статистика застосовує метод оцінки загальної сукупності на основі вивчених окремих одиниць або на основі вибіркових сукупностей.

**Статистичні висновки.** Статистичні висновки про властивості генеральних сукупностей за вибірковими завжди мають ймовірнісний характер, тобто робляться з певним ступенем безпомилковості і ніколи не робляться з повною достовірністю.

Статистичні висновки, як головна складова частина методу дослідження масових явищ, мають свої відмінні риси: їх роблять з чисельно вираженою визначеністю. Теоретичною основою для їх побудови є розділ математики, що вивчає закономірності випадкових подій – теорія ймовірностей. Водночас, факт, що результати статистичного спостереження відібрані у випадковому порядку з відповідних генеральних сукупностей, дає можливість відповідно до теорії ймовірностей оцінити ступінь відхилення результатів спостереження від відповідних показників генеральної сукупності. Таким чином, ймовірнісна основа варіаційної статистики дозволяє оцінити ступінь точності одержуваних результатів досліду. Основу

вивчення природних процесів становить виявлення причинно-наслідкових зв'язків між явищами експериментальним шляхом.

**Теорія ймовірностей**. Здійснивши за своїм бажанням одне або кілька початкових явищ (надалі вони називаються чинниками), експериментатор отримує можливість вивчати явища, що з'являються – наслідки. Іноді в процесі експерименту вдається зробити випадкове відкриття, тобто виявити явище-наслідок, про який раніше нічого не було відомо. Але, як правило, експериментатор заздалегідь намічає явища-наслідки, появу яких він очікує. При цьому, найскладніше явище можна розбити на дрібніші явища, щодо яких залишається з'ясувати: відбулися вони або ні. Наприклад, обробляючи насіння на схожість певним препаратом, експериментатор міг поставити задачу оцінити ефект різних його доз. У якості ефекту можна прийняти число пророслого чи не пророслого насіння. Вимірюючи масу будь-якої речовини, як окремих локальних явищ можна розглядати всілякі апріорні значення цієї маси. Завдання експериментатора, таким чином, зводиться до спостереження того, які зі значень маси здійснилися, а які – ні.

Явища, що розглядаються з цієї точки зору, тобто здійснилися вони чи ні, називаються **подіями**. Відносно подій ставиться основне завдання: передбачити чи з'явиться досліджувана подія під час здійснення деякого наперед заданого комплексу чинників (явищ – причин). Подія, яка при заданому комплексі чинників обов'язково станеться називається **достовірною, а подія**, яка не може статися, називається **неможливою подією.** Судження про достовірність або неможливість деякої події є категоричними судженнями. Такі судження прийнято вважати остаточним результатом дослідження. Звідси виникає інтерес до зворотній задачі: вказати комплекси чинників при яких про задану подію можна зробити такі категоричні судження.

Однак кожна подія – результат дії багатьох чинників, частину з яких іноді можна передбачити або організувати у досліді. У такому випадку категоричне судження про подію є неможливим. Виходить наступна ситуація: задані чинники сприяють події, і, отже, вона може відбутися. З іншого боку, лише дії цих чинників недостатньо, щоб гарантувати появу події, і, значить, вона може і не відбутися. Подія, яке при заданому комплексі чинників може або відбутися, або не відбутися, називається **випадковою подією.** Наприклад, досліджується врожайність культур. Такі чинники, як технологія обробітку, внесення різних доз добрив і т.д. можна організувати у досліді, тобто врахувати – ці чинники є основними. Інша група чинників є невідомою або не піддається обліку. Ці чинники при статистичному аналізі отримали назву **випадкових.**

Для того щоб з'ясувати, відбудеться або не відбудеться певна подія при заданому комплексі чинників, потрібно здійснити цей комплекс, тобто провести **випробування. Випробуванням** є будь-який експеримент, в результаті якого проводять спостереження. Передбачити результат одиничного випробування можна тільки для достовірних або неможливих подій, оскільки випадковість події не модна оцінити одиничним

випробуванням. Будь-яка випадкова подія за одиничного випробування було б оцінена як достовірна, якщо події є, а як неможлива – якщо події не відбулися. Теорія оцінки випадкових подій базується на великому числі випробувань, тобто для масових подій.

Важливою умовою при цьому є незмінність комплексу основних чинників. Події, що відбуваються при одному і тому ж комплексі чинників, називаються **однорідними.** Встановлено, що однорідні випадкові події у великій їх масі підкоряються деяким закономірностям. Ці закономірності отримали назву **ймовірнісних**. Характер імовірнісних закономірностей можна усвідомити на наступному прикладі. *Приклад.* Під час підкидання монети можливі дві події: випадання монети гербом або решкою. Події з однаковими можливостями здійснення називаються **рівно можливими.** Так, при симетричній монеті випадання герба і цифри – **рівно можливе.** Однак, якщо б було здійснено, наприклад, 1000 підкидань, і з них 700 раз випав герб, то для наступної серії випробувань можна було б прогнозувати, що герб з'явиться в 70 % випадків. Причому таке відхилення від очікуваних 700 появ герба з 1000 підкидань можна було б вважати пов'язаним з несиметричністю монети.

Встановлене у результаті досліду відношення числа появи події до загального числа всіх випробувань називається **частотою події.** У зазначеному прикладі з монетою частота випадання герба дорівнює 0,7.

З прикладу можна зробити висновок, що частота події, яка виступає як деяка статистична закономірність, є пов'язаною з внутрішніми характеристиками події. Частота є мірою цих внутрішніх характеристик події. Вона тим надійніша, чим більше число випробувань було здійснено. При дуже великому числі випробувань частота майже перестає змінюватися, наближаючись до деякої величини. Цю величину і можна прийняти за нестандартну для нас числову характеристику. Так, під час підкидання монети 4, 12 і 24 тис. раз частота появи герба відповідно дорівнювала 0,6080; 0,5016; 0,5005. Очевидно, що вона тут наближається до числа 0,5. Числова характеристика випадкової події, що володіє тією властивістю, що для будь-якої досить великої серії випробувань частота події лише незначно відрізняється від цієї характеристики, називається **ймовірністю події.** З огляду на це, можна встановити, що ймовірність є тією теоретичною межею, до якого прагне частота подій при збільшенні числа випробувань. Ймовірність – ідеальне вираження частоти подій. Таке визначення ймовірності називається **статистичним.** Це визначення не є досить жорстким з точки зору математики. За статистичним визначенням важко вивчати властивості ймовірності. Однак є і ряд позитивних його властивостей. Так, статистичний підхід дозволяє знаходити ймовірності подій, структура яких невідома. Наприклад, тільки статистичний підхід дозволив визначити ймовірність народження хлопчиків, рівну 0,52 і дівчаток – 0,48. Водночас, існують два інших, зручніших з формальної точки зору, визначення ймовірності: класичне і геометричне. Однак для них потрібно знати структуру розглянутих подій. Поняття про геометричне визначенні

ймовірності можна отримати з таких прикладів. *Приклад 1.* Припустимо, що в деякому квадраті випадковим чином вибирається точка. Яка ймовірність, що вона виявиться в області D? Очевидно, що ймовірність ця буде тим більшою, чим більшою є область D. В якості визначника ймовірності тут виступає площа. Ймовірність того, що випадкова точка потрапить в область D (здійснення події D) дорівнює: p (D) = SD / S, де SD – площа області D; S – площа всього квадрата. Так, геометричне визначення ймовірності допустиме не тільки для площини, а й для прямої або простору. У першому випадку основою для визначення ймовірності служить певний відрізок, а випадковим подіям відповідають його частини. Ймовірність обчислюється як відношення довжини частин до загальної довжини відрізка. У другому, випадку основою до визначення приймають деякий куб, випадковим подіям відповідають різні тіла, розташовані в кубі. Ймовірність обчислюють як відношення обсягів тіл до обсягу куба.

Найбільший інтерес представляє класичне визначення ймовірності. З цим визначенням пов'язані основні теореми теорії ймовірностей. Ймовірність тут визначається апріорі, тобто до визначень, виходячи з певної структури випадкових подій, а саме з розбивки на рівно можливі наслідки. *Приклад 2.* Нехай при підкиданні монети поява герба або цифри будуть досліджуваними подіями a й b. Причому, якщо при одному киданні відбудеться подія a, то не відбудеться іншої події b. Такі події називають **несумісними.** Кожну з подій називають результатом випробування. Враховуючи рівно можливість випадків у випробуванні, ймовірність кожної події є рівною. При одиничному киданні кубика з 6 гранями (наприклад, 1, 2, 3, 4, 5, 6 очок), ймовірність появи будь-якої однієї грані p = 1/6.

Результати випробування є найпростішими випадковими подіями. Можна розглядати складніші події, які об'єднують кілька випадків. Наприклад, при киданні грального кубика ми можемо очікувати таку подію, як випадання числа очок більше 2. У такому випадку говорять, що появі події з випаданням більше двох очок, тобто з 3, 4, 5 і 6 очками, сприяють чотири результати з шести. Імовірність цієї події p = 4/6. Таким чином, визначається класичне визначення ймовірності. **Ймовірністю випадкової події** називається відношення числа ходів, що сприяють події, до числа усіх можливих результатів. Отже, коли ймовірності незалежних подій відомі апріорі, то можна визначити ймовірні чисельності будь-якого даного числа n, n1, n2 .... прояву чи не прояву події. При цьому неважливо, рівними або нерівними є p і q, головне, щоб вони залишалися при випробуваннях постійними. Цей факт має велике значення у теорії статистики і біометрії.

При вивченні природних явищ виділення елементарних подій і взагалі розчленування причинного процесу, в результаті якого відбуваються випадкові події, зазвичай неможливо. Класичний підхід до визначення ймовірності тут безсилий. Проблему визначення ймовірностей таких подій вирішують на основі статистичного підходу. Однак класичний підхід до визначення ймовірностей подій лежить в основі теорії аналізу випадкових подій і теоретичних (модельних) розподілів результатів випробувань.

## *Лекція 3. Описова статистика*

**Характеристика сукупності.** Будь-яка множина окремих відмінних один від одного і в той же час подібних в деяких істотних відносинах, об'єктів становить так звану **сукупність** (популяції рудих полівок того чи іншого району, стадо корів даного господарства, потомство певного бика, заготовлювані в області або краї білячі шкурки, рослини на дослідних ділянках, група курчат, на яких ставиться дослід щодо застосування антибіотиків, мальки окуня в озері і т. д.). Поняття сукупності можна застосувати не тільки до тварин і рослин. Такими ж сукупностями є, наприклад, діти, що народилися в країні протягом якогось року чи місяці або молекули газу в тому чи іншому об'ємі. До складу сукупності входять різні члени або одиниці: для популяції тварин – кожна окрема тварина, для стада корів одиницею є кожна корова, для сукупності шкурок – кожна шкурка, для потомства бика – кожне отримане від нього теля, для сукупності зерен гречки – кожне окреме зерно.

Зазвичай число одиниць сукупності називають обсягом сукупності і позначають латинською буквою *n*. Одиниця сукупності може характеризуватися певними ознаками, наприклад: корови – надоями за лактацію, вагою, мастю; молекули газу – швидкостями їх руху і т. д. Кожна досліджувана ознака приймає різні значення у різних одиницях сукупності, вона змінюється в своєму значенні від однієї одиниці сукупності до іншої. Ця різниця між одиницями сукупності називається **варіацією або дисперсією** (розсіюванням). Коли говорять, що ознака варіює, це означає, що вона приймає різні значення сукупностей. Так, ознака у корів даної породи, мишей дослідної групи, поросят одного посліду і т. д. Значення або міру ознаки одиниці сукупності називають **варіантом** і позначають літерою *X*. Значок *i* – порядковий номер варіанту. Незважаючи на відмінності між варіантами за значенням досліджуваної ознаки, сукупність цих варіантів володіє однорідністю. Шерсть вівці неоднакова за забарвленням, розміром, якістю хутра, але вона однорідна, так як це – шерсть особин одного і того ж виду – вівці звичайної.

**Розрізняють сукупності:**

1. Генеральну
2. Вибіркову (вибірка).

**Генеральна** – теоретично нескінченна сукупність усіх одиниць або членів, які можуть бути віднесені до неї. Через нескінченно велике число членів, генеральну сукупність вивчити практично неможливо. Тому з неї вибирають частину для безпосереднього вивчення, тобто вибірку.

**Існує кілька способів відбору варіантів у вибірку:**

1. плановий відбір (груповий відбір; гніздовий (або серійний));
2. стихійний відбір (механічний).

Єдиною стабільною умовою є однорідність відібраного у вибірку матеріалу. Завданням вивчення будь-якої сукупності є отримання

статистичних (або, як іноді кажуть, біометричних) характеристик, або показників, які дозволяють судити про дану сукупність в цілому, про відмінності всередині неї і про відмінність її від інших, схожих з нею або близьких до неї сукупностей. Сукупність стає статистичною тоді, коли її опис доповнюється кількісним методом. Застосування кількісного методу вивчення сукупностей дозволяє отримувати для неї статистичні характеристики, за допомогою яких отримують основну інформацію про сукупності.

**Варіюючі ознаки та їх облік.** При вивченні одиниць сукупності за ознакою необхідно записати отримані дані і згрупувати їх. Способи згруповання залежать від характеру варіації досліджуваних ознак.

Розрізняють такі **типи варіації ознак:**

- якісна;
- кількісна

Якщо відмінності між варіантами виражаються в якихось якостях, то таку варіацію називають **якісною.** Якщо сукупність тварин характеризують по масті, тоді кожен варіант повинен отримати якісну характеристику відповідно до заздалегідь прийнятих позначень: чорна, руда, чорно-ряба, біла і т. д. У цьому простому випадку підрахунок числа особин в кожній з виділених груп дає уявлення про склад популяції в цілому.

В інших випадках відмінності між варіантами будуть **кількісними.** Кількісна варіація може бути двох типів: **переривчаста (дискретна) і безперервна**. У першому випадку відмінності між варіантами, окремими значеннями випадкової змінної, виражаються цілими числами, між якими немає і не може бути переходів. Наприклад, кількість дитинчат (поросят у свиноматок, лисенят у сріблясто-чорних лисиць), число променів у плавцях риб, кількість пелюсток у квітці, число хребців у птахів і т. д. Для вивчення подібного варіювання треба порахувати у кожної одиниці сукупності число досліджуваних елементів і записати його на відповідну картку. При безперервній варіації значення варіантів не обов'язково виражається тільки цілими числами. Все залежить від того, який ступінь точності приймається для характеристики даної кількісної ознаки. Так, наприклад, при вивченні ваги великої рогатої худоби можна обмежитися значеннями варіантів, вираженими в кілограмах, відкинувши грами, але зовсім недостатньо округляти до кілограмів вагу риб, так як грам тут має велике значення. У дослідах ж по вивченню впливу гормонів на ріст гребенів у курчат вагу гребеня доведеться вимірювати у міліграмах. Молочну продуктивність під час лактації зазвичай виражають у кілограмах, але загальна картина надоїв не зміниться, якщо округляти її до десятків кілограмів. Оцінка ж жирності молока – у відсотках, що виражені цілими числами, явно недостатня, її треба подавати з урахуванням десятих і навіть сотих часток відсотка. Однак у всіх цих і їм подібних випадках існує безперервна варіація, що виражається в тому, що між варіантами можливі всі переходи. При вивченні безперервної варіації треба усі одиниці сукупності характеризувати кількісно з тим

ступенем точності, яка заздалегідь намічена і найбільше підходить у даному конкретному випадку.

**Згрупування даних при якісній варіації.** Щоб проаналізувати ту чи іншу сукупність, необхідно згрупувати отримані окремі варіанти і потім представити це згрупування у вигляді таблиці або ряду. При упорядкуванні отриманих даних легко обробити їх математично і вивести статистичні показники, які будуть вичерпно характеризувати досліджувану сукупність. Проблема згрупування займає велике місце у статистиці взагалі, так як помилкове згрупування даних може призвести до неправильних висновків про суть досліджуваного явища. Так, якщо норки розрізняються за забарвленням, то їх розподіл може бути виражений у кількості тварин кожного забарвлення і в процентах, які складають норки кожної забарвлення від загальної кількості тварин.

Окремим випадком якісної варіації є **альтернативна**, коли в сукупності можна виділити тільки дві групи. У членів однієї групи присутня певна якість (або ознака), у членів іншої групи її немає. Так, при перевірці на туберкульоз тварини розпадаються на 2 групи – з позитивною реакцією і з негативною. Одні корови в даному стаді рогаті, інші – комолі і т. д. Водночас, при кількісній варіації необхідно попередньо намітити для таблиці класи, які охоплюють всі отримані кількісні дані від мінімальних до максимальних. Це доцільно робити при переривчастій (дискретній) кількісній мінливості.

**Варіаційний ряд і його графічне зображення.** Після розподілу варіантів за класами виходять ряди, що показують як часто зустрічаються варіанти кожного класу і як варіює ознака від мінімуму до максимуму. **Варіаційний ряд** – подвійний ряд чисел, що показує розподіл варіантів за їх частотою або чисельності зустрічання. За таким рядом можна судити не тільки про межі, а й про характер варіації. Клас, що володіє найбільшою частотою, отримав назву **модального**, значення ж крайніх класів називають **лімітами або межами.**

Будь-який варіаційний ряд можна зобразити графічно. Графічне зображення варіаційного ряду у загальному вигляді отримало назву **кривої розподілу або варіаційної кривої.** Існують два способи графічного зображення конкретних варіаційних рядів. Перший з них застосовується при дискретній варіації, але в тому випадку, коли класи намічені за окремими значеннями варіантів і носить назву **полігон розподілу**. На осі абсцис нанесені класи, на осі ординат – частоти. Висота кожного класу, що пропорційна його частоті відмічається колом. При безперервній варіації, якщо класи намічені на межових осях, на осі абсцис наносять нижні межі класів, на осі ординат – частоти. Такий графік носить назву – **гістограми.**

Під час статистичної обробки матеріалу виникає питання: скільки класів необхідно намічати? Це залежить від: обсягу сукупності та від величини варіаційного розмаху. На практиці можна керуватися приблизно такими правилами:

| Кількість варіант | Число класів |
|---|---|
| 25-40 | 5-6 |
| 40-60 | 6-8 |
| 60-100 | 7-10 |
| 100-200 | 8-12 |
| більше 200 | 10-15 |

Варіаційний ряд безперервної мінливості також може бути зображений на графіку. У такому випадку необхідно будувати гістограму, тобто ступінчасту діаграму.

**Характер розподілу варіантів у варіаційному ряду.** Вивчаючи розподіл варіантів у варіаційних рядах легко помітити деякі загальні закономірності, а саме:

1. більшість варіантів розташовується у середній частині варіаційного ряду або десь у середині варіаційної кривої, тут спостерігається максимум варіантів, іншими словами їх скупчення;
2. розподіл варіантів по обидві сторони від цього максимуму є приблизно симетричним;
3. частота варіантів поступово зменшується у напрямку до меж варіаційного ряду.

Ці закономірності тією чи іншою мірі властиві будь-якому варіаційному ряду, оскільки його закономірності ґрунтуються на закономірностях випадкової варіації, що вивчаються теорією ймовірностей.

**Групи показників для характеристики варіаційних рядів.** У попередній лекції ми розглянули способи зведення даних, що становлять статистичні сукупності у варіаційні ряди. Кожен варіаційний ряд і його графічне зображення – це ніби «скупчення» вихідного фактичного матеріалу та перетворення його у наочну форму. Однак, зазвичай, цього недостатньо, тому дуже важливим є отримання характеристик для сукупності, які були б виражені цифровими показниками. З їх допомогою можна було б порівнювати різні ряди. Одним з найпростіших способів кількісної характеристики варіаційного ряду є вказівка на його розмах, тобто на верхню і нижню межі, які зазвичай називають **лімітами**. Якщо, наприклад, відомо, що варіаційний ряд за молочною продуктивністю одного стада корів має розмах від 2000 до 4000 кг, а іншого – від 2500 до 6800 кг, то, здавалося б, можна зробити висновок про вищу якість другого стада. Однак ліміти не вказують на те, як саме розподіляються за визначеною ознакою окремі члени сукупності. Тому для характеристики сукупності потрібні ще такі показники, які відображали б властивості усіх її членів.

**Варіаційні ряди можуть відрізнятися:**

- за тим значенням ознаки, навколо якої концентрується більшість варіантів. Це значення ознаки відображає рівень її розвитку у даній сукупності (центральну тенденцію ряду) тобто, типове для ряду;
- за ступенем варіації варіантів навколо рівня, тобто за ступенем відхилення від центральної тенденції ряду.

Відповідно до цього статистичні показники поділяються на **дві групи:**

- показники, які характеризують центральну тенденцію або рівень ряду,
- показники, що вимірюють власне ступінь варіації.

**До першої групи** належать різні середні величини: мода, медіана, середнє арифметичне, середнє геометричне. **До другої** – варіаційний розмах, середнє абсолютне відхилення, середнє квадратичне відхилення, дисперсія, коефіцієнти асиметрії та варіації. Існують ще й інші показники, але у біологічній статистиці вони застосовуються вкрай рідко.

***Мода і медіана.*** Під час вивчення розподілу самок собаки за числом цуценят виявилося, що 39 самок із загального числа 80 мали по 4 цуценяти, тобто клас «4 цуценяти» володів найбільшою частотою. Такий клас називається **модальним.** Значення ж модального класу називають **модою.** Мода позначається символом *Mo.* Величина моди є типовою для усієї сукупності. Дійсно, у даному прикладі майже половина самок з 80 мала у посліді саме 4 цуценяти.

Для ряду розподілу змій за кількістю хвостових щитків (табл. 1) модальним є клас «46-48 щитків». А так як клас тут охоплює кілька значень варіантів, то для його характеристики треба додатково потрібно обчислити середнє значення класу. Воно дорівнює 46 + 48/2 = 47. У такому випадку *Mo* = 47 щитків. До числа середніх величин відноситься також **медіана**. **Медіана** – це значення варіанту, що знаходиться точно у середині ряду (позначається *Me*). Щоб знайти такий варіант, треба спочатку розташувати усі варіанти впорядковано від мінімальних їх значень до максимальних. Таке розташування варіантів називають **ранжуванням**. Щоб визначити *Me* за парного числа варіантів, потрібно взяти значення двох сусідніх серединних варіантів (наприклад, при n=80 значення варіантів із порядковими номерами 40 і 41), і розділити їх суму на 2. У прикладі, представленому вище, обидві ці варіанти будуть мати значення «4 цуценяти», отже, *Me* даного ряду = 40.

Медіана і мода дають чітке уявлення про сукупність в цілому. Вони характеризують свого роду типовий розмах у даній сукупності (звичайно, мова йде тільки про певну ознаку). Використання моди і медіани у біології на сьогодні досить обмежене, але в деяких випадках без них дуже важко обійтися, зокрема, якщо отримані дані не є чисто кількісними, а тому не можуть бути представлені у вигляді точного варіаційного ряду. Так, наприклад, тяжкість захворювання піддослідних тварин або їх вгодованість можна умовно оцінювати ступенями: слабка, задовільна, середня, висока; або балами 1, 2, 3 і т.д. Тоді мода чи медіана можуть досить добре характеризувати типовий розмах у сукупності. Проте, у випадках коли вивчається досить однорідна сукупність, а варіація всередині неї є чисто кількісною, вигідніше користуватися іншими середніми величинами.

***Середнє арифметичне та його властивості.*** Знаходження середнього арифметичного – це по суті заміна індивідуальних варіюючих значень ознак окремих членів сукупності деякою однаковою величиною при збереженні основних властивостей усіх членів сукупності. Цій умові найбільшою мірою

задовольняє таке поняття як **«середнє арифметичне»**, що позначається – *X* (у літературі інколи зустрічається позначення *M*).

*Приклад.* Ряд членів сукупності, тобто ряд значень випадкової змінної *x1, x2 ... xi* замінимо таким же рядом з однакових величин *x*, тобто x, x, x … x (n разів). Тоді сума усіх варіантів сукупності *x1 + x2 + x3 + ... + xn* буде дорівнювати *X + X + X + ... + X* (n разів), тобто *nX.* Суму усіх варіантів сукупності можна скорочено позначити *x*, (*x* – позначає значення кожного варіанту; грецька буква *Σ* – велика сігма – позначає суму; суми часто позначають також латинською літерою *S*). Тоді ***Σ x, = nx,*** звідки ***X = Σ x / n***.

Ми отримали найзагальнішу і, водночас, найпростішу формулу середнього арифметичного. Для того щоб обчислити середнє арифметичне, досить скласти значення всіх варіантів і суму розділити на загальне число варіантів. Власне, у найпростіших випадках так проводять розрахунки. Очевидно, у таких випадках можна користуватися даними, отриманими безпосередньо під час аналізу членів сукупності, не вдаючись до угруповання варіантів.

Однак при великій кількості варіантів цей прямий спосіб визначення середнього арифметичного за вказаними формулами виявляється не таким зручним. Крім того, при його застосуванні неможливо розрахувати деякі інші біометричні показники. Тому на практиці часто користуються обхідними методами обчислення середнього арифметичного на основі вже згрупованих даних. Ці методи будуть розібрані пізніше.

**Властивості середнього арифметичного:**

1. Якщо кожен із варіантів сукупності, для якого обчислюється середнє арифметичне, збільшити або зменшити на одну і ту ж величину, то і середнє арифметичнє відповідно збільшиться чи зменшиться на стільки ж.

*X1 / A; X2 / A; X3 / A; X4 / A; Xi / A; то Σxi / A*

*2.* Алгебраїчна сума відхилень окремих варіантів від середнього арифметичного (тобто відмін між кожним конкретним значенням ознаки і середнього арифметичного) рівна нулю:

*Σ (xi - X) = 0.*

*3.* Сума квадратів відхилень від середнього арифметичного буде менше суми квадратів відхилень від будь-якої іншої величини А не рівної Х, тобто:

*Σ (xi-X) 2 <Σ (xi-A) 2,* якщо А не дорівнює Х.

**Значення середнього арифметичного та його сутність.** Середнє арифметичне, як і деякі інші середні, відоме давно. Немає буквально жодної біологічної роботи, в якій не зустрічалися б в тій чи іншій формі середні арифметичні. Воно є узагальнюючою величиною, яка конденсує у себе всі особливості даної сукупності або ряду розподілу. Вона відображає рівень всієї сукупності в цілому, дає зведену, узагальнену характеристику даної досліджуваної ознаки.

Цифрове значення середнього арифметичного як таке може не зустрітися ні в одному конкретному випадку в сукупності. Може виявитися, що жоден варіант не буде їй рівний. Наприклад, якщо середнє число курчат у сріблясто-чорних курей рівне 4,7, то, очевидно, фактичне число курчат ніяк

не може бути дробовим. У такому випадку, середнє арифметичне є абстрактною величиною. Водночас, ця величина виражається в тих же одиницях виміру, що і варіанти ряду. При визначенні середнього арифметичного відкидаються випадкові коливання, відхилення від центральної тенденції, від рівня варіаційного ряду і виступає загальний закон явища. Розкривається типовий розмах для всієї сукупності в цілому. У той же час потрібно враховувати **можливі помилки** у розумінні середнього арифметичного:

- Середнє арифметичне характеризує всю сукупність в цілому, а не окремі члени сукупності. Так, середнє число цуценят – 4,7 відноситься тільки до всієї групи, кожна ж окрема собака характеризується своїм числом цуценят – від 1 до 9.
- Середнє має сенс тільки відносно якісно однорідної сукупності. Так, не можна обчислювати середню вагу тварин для групи, що включає і молодняк різного віку, і дорослих тварин. Треба враховувати кожну вікову групу окремо і для них обчислити *X.*
- Оскільки середнє арифметичне відноситься до даної сукупності, перенесення її на явища, що виходять за її рамки є ризикованим без спеціального аналізу питання про правомірність такого перенесення.
- Середнє відноситься лише до окремих досліджуваних ознак і не може бути автоматично перенесене на їх суму.

**Вимірювання варіації: варіаційний розмах і середні відхилення.** Середнє арифметичне вказує на те, яке значення ознаки є найхарактернішим для даної сукупності. Але само по собі значення цієї величини ще недостатнє для характеристики сукупності, так як головною особливістю сукупності є наявність різноманіття між її членами, тобто варіації. Якби не було варіації, то інформацію про сукупності можна було б отримати лише за одним її членом. При наявності ж варіації ця інформація повинна бути включати характер і ступінь варіації.

Облік варіації тієї чи іншої ознаки у сукупності має дуже велике значення для біолога, так як будь-яка варіація у популяції тварин чи рослин у кінцевому рахунку відображає відмінності між організмами – в їх спадковій природі і в тих умовах, при яких вони зростали чи вирощувалися. Прийоми роботи з тваринами повинні змінюватися залежно від характеру їх варіації. Без оцінки варіації неможливо і порівняння двох сукупностей. Так, два стада корів можуть мати дуже близькі середні надої, але в одному стаді величини надоїв сильно різняться, в іншому ж – корови формують досить однорідну групу з невеликим розмахом коливань. Визначення варіаційного розмаху або різниці між максимальним і мінімальним значеннями варіантів, може певною мірою вказувати на ступінь варіації, але цього недостатньо. По-перше, крайні величини у рядах є не дуже стійкими, і за умови зміни кількості досліджуваних особин вони легко зсуваються. По-друге, за одних і тих же меж варіації розподіл варіантів у рядах може бути різним. Саме тому, для характеристики відмінностей між окремими значеннями випадкової змінної *x* (варіації між членами сукупності), потрібен такий показник, який

узагальнював би коливання усіх варіантів. Для цього потрібно порівнювати варіанти або один з одним, або з якоюсь однією постійною величиною. У якості останньої найкраще взяти середнє арифметичне.

**Дисперсія і середньоквадратичне відхилення.** Досконалішим показником, що характеризує варіацію, є середній квадрат відхилень варіантів від середнього арифметичного, що має назву **дисперсія**, і середньоквадратичне відхилення, або **стандартне відхилення.** Дисперсію позначають *σ2* (грецька літера сігма) або *s2* (латинська літера ес), а середньоквадратичне відхилення – *σ*. В цілому, це можна формулювати так: **дисперсія** – це сума відхилень окремих значень варіантів від середнього арифметичного, поділена на загальну кількість варіантів, а **середньоквадратичне відхилення** – корінь квадратний з цього локального значення дисперсії. Незважаючи на те, що після вилучення кореня квадратного виходять значення зі знаками плюс і мінус, зазвичай беруть тільки позитивне значення.

**Ступені свободи.** Величина *n – 1* отримала особливу назву – **число ступенів свободи (**точніше, **число ступенів свободи варіації)**. ЇЇ позначають літерами *df*.

Існують різні способи обчислення статистичних показників:

- прямий через значення варіантів (без варіаційного ряду, при малому *n*);
- прямий через значення варіантів для варіаційного ряду;
- непрямий метод (умовного середнього).

Отже, для визначення статистичних показників потрібно провести досить велику обчислювальну роботу, але її обсяг може бути скорочений правильним вибором методу, що найбільше підходящого для обробки даного матеріалу і застосуванням наявних технічних засобів Найкраще користуватися прямим способом обчислень, так як він дає найточніші результати. Непрямий спосіб через штучну розбивку матеріалу на класи завжди супроводжує відома неточність.

**Середнє геометричне.** Середнє арифметичне – однозначно найчастіше застосовуваний статистичний показник, у тому числі у біології. Однак, у деяких випадках (наприклад, при вивченні темпів зростання організмів або зростання цілих популяцій) доводиться користуватися іншою середньою величиною – середнім геометричним.

Формула для її обчислення наступна:

$$X\ g = \sqrt{}\ x1\ .x2\ .x\ n = \sqrt{}\ \Pi\ xi$$

Очевидно, що при її визначенні треба виключати варіанти, що виражаються нулем або негативним числом. На практиці обчислення середнього геометричного проводиться за допомогою логарифмів за такою робочої формулою:

$$\text{Log } X\ g = 1 / n\ (\log x1 + \log x2 + \log xn),$$

тобто, логарифм середнього геометричного дорівнює арифметичному середньому суми логарифмів окремих значень *x*. За значенням *log X* потім визначається величина *x*.

## *Лекція 4. Статистична гіпотеза. Вибірковий метод та репрезентативність вибіркових даних*

**Проблема достовірності у статистиці**. Прийоми і методи, викладені у попередніх лекціях дають можливість вичерпно охарактеризувати біологічні сукупності. Кожна сукупність може бути представлена у вигляді ряду розподілу. Для ряду розподілу можна визначити статистичні показники, що вказують на найтиповіший рівень розвитку досліджуваної у сукупності ознаки і на ступінь варіації окремих одиниць сукупності навколо цього рівня. Більшість з них – іменовані величини (середнє арифметичне, мода, медіана, середнє відхилення), деякі виражаються у відсотках (коефіцієнт варіації) або є неіменованими числами (дисперсія, коефіцієнт асиметрії). Але, враховуючи те, що вони – статистичні величини, а, отже, засновані на вивченні масових явищ, виникає дуже важливе теоретичне і практичне питання про те, наскільки вони є достовірними.

Проблема достовірності займає чільне місце у статистичній теорії. Генеральна сукупність – це все сукупність даних об'єкту, що підлягає вивченню. Вона може розглядатися як така, що складається з нескінченно великої кількості окремих одиниць, а та частина об'єктів, яка піддається дослідженню, називається **вибірковою сукупністю або просто вибіркою.**

Обидва типи сукупностей у загальному характеризуються однаковими закономірностями випадкової варіації. Для їх характеристик можуть бути обчислені статистичні показники: середнє арифметичне і середнє квадратичне відхилення. Оскільки середнє арифметичне позначається символом $x$, то нехай тепер $x$ позначає середнє арифметичне вибіркової сукупності. Середнє арифметичне генеральної сукупності позначаємо $\mu$. Яким же тоді є співвідношення між $x$ і $\mu$? Припустимо, що для сукупності, що складається зі 168 корів симентальської породи, було отримане середнє арифметичне глибини грудей 73,8 см. 168 корів є вибіркою з генеральної сукупності, яка охоплює популяцію усіх корів симентальської породи. Якби було взято ряд вибору з популяції симентальської породи, то виявилося б, що $x$ цих вибірок будуть різними. Одні з $x$ будуть дещо більшими ніж 73,8 см, інші – меншими. Дуже важливим є те, що розподіл вибіркових середніх при достатній їх кількості близький до нормального, тому до нього ставляться стандартні закономірності. Виявляється, що окремі значення середніх арифметичних вибірок $(x)$ варіюють навколо середнього арифметичного генеральної сукупності. Варіація ж вибіркових середніх навколо $\mu$ може бути визначена своїм середнім квадратичним відхиленням або своєю сігмою. Така сігма отримала назву **середньої помилки** або **середньої квадратичної помилки (**іноді її називають також **стандартною помилкою M)**. Саме вона вказує на ступінь близькості $x$ і $\mu$.

**Середня помилка** M для $x$ може бути обчислена за формулою:

$$M*x = \delta / \sqrt{n}$$

У знаменнику формули під коренем *n* – обсяг вибіркової сукупності. Це означає, що величина середньої помилки обернено пропорційна чисельності вибіркової сукупності. У прикладі з глибиною грудей у симентальських корів відомо, що n = 168 і δ = 2,45. Звідси середня помилка для середньої арифметичної глибини грудей вивчених 168 симентальських корів: δ = 2,45 / 168 = 0,17.

**Середня помилка – помилка вибірковості.** Термін *«помилка»* часто вводить в оману початківців, які припускають, що вона є результатом недостатньої акуратності у роботі. Однак, це не є вірним. **Середня помилка – це статистична помилка**. Вона не має нічого спільного із помилкою точності. Логічно, що всі вимірювання (ваги і промірів риб, надоїв корів і жирності їх молока, настриг вовни овець та її довжини тощо) треба робити точно і сумлінно. Але статистичні показники для вибіркової сукупності завжди мають так звані **помилки вибірковості** (їх також називають **помилками репрезентативності**), які являють собою середню величину розбіжності між середніми значеннями досліджуваних ознак у вибірках і генеральній сукупності. Так як $mx = \delta / \sqrt{n}$, то, очевидно, що розмір визначеної середньої помилки залежить від сігми вибіркової популяції та від її обсягу. Чим краще взята вибірка і, чим більше її розміри, тим менше і середня помилка, і тим меншим буде розходження між значеннями ознак у вибіркових і генеральної сумах.

Біолог майже завжди має справу з вибірками – і при проведенні дослідів з тваринами або рослинами, а також при вивченні матеріалу, взятого з природи; при цьому генеральні сукупності залишаються невідомими. Тому вчений повинен постійно пам'ятати про той ризик, який супроводжує його висновки. Часто ці висновки ґрунтуються на вивченні невеликого матеріалу, тому отримані у дослідах або спостереженнях статистичні показники можуть мати значні статистичні помилки. Легко помітити, що внаслідок коливання вибіркових середніх навколо середньої генеральної сукупності один окремий досвід може дати результат, що відхиляється від істинного на 2 або навіть 3 помилки. Але при значній кількості дослідів їх результати будуть групуватися близько до центру розподілу генеральної сукупності, тобто до $\mu$, що дає можливість впевнено зробити правильний висновок.

Деяка похибка органічно буде притаманна результатами спостереження, проведеного на основі вибірки. Цю похибку і вимірює середня помилка, яка тому й називається **помилкою вибірковості**. Разом з тим, абсолютно необхідно, щоб вибіркова сукупність досить добре відображала генеральну сукупність, інакше судження про генеральну сукупність за вибіркою буде неправильним, незважаючи на правильність статистичних обчислень. Домогтися правильного відображення генеральної сукупності можна за однієї неодмінної умови – відборі варіантів для вибірки на основі випадковості. Чим більшою мірою цей відбір буде випадковим, тим правильнішими будуть висновки, що робляться на основі вибіркової сукупності. Саме тоді можна покладатися на результати вибіркового спостереження.

Найпростіший спосіб отримання випадкових вибірок – відбирають екземпляри за допомогою таблиці випадкових чисел. На принципі випадковості ґрунтуються різні схеми відбору варіантів для вибірки: **випадкова безповторна вибірка**, коли взяті для вибірки варіанти вже не повертаються назад у генеральну сукупність, **випадкова повторна вибірка** з поверненням взятих для вибірки варіантів назад у генеральну сукупність і т.д.

**Закон великих чисел.** У зв'язках між статистичними показниками вибіркових і генеральних сукупностей виражається так званий **закон великих чисел**. У найзагальнішому вигляді цей закон полягає в тому, що чим більшим є число $n$ деяких випадкових величин, тим їх середнє арифметичне буде ближчим до середнього арифметичного генеральної сукупності, і тим меншою буде різниця між $x$ і $\mu$. По факту збільшення $n$ ймовірність наближення $x$ до $\mu$ стає все більшою, прагнучи при цьому $n = \infty$ до одиниці, тобто до повної достовірності. У цьому полягає теорема одного з основоположників математичної статистики російського математика Пафнутія Львовича Чебишева. Оскільки будь-яке явище, як правило, складається з маси одиничних, випадкових явищ, то закон великих чисел виступає як реальний закон об'єктивної дійсності. Саме він лежить в основі нормального розподілу варіантів у варіаційному ряду, тобто розподілу значень випадкової змінної $x_i$ навколо $X$, а також в основі розподілу вибіркових $X$ навколо μ.

Вибіркові середні для яких обчислюються середні помилки, є такими ж випадковими величинами, як і значення варіантів у звичайному варіаційному ряду. Із зростанням обсягів вибірок їх варіація навколо генеральної середньої стає все меншою. Середнє ж арифметичне з усіх вибіркових середніх дорівнює середньому арифметичному генеральної сукупності, тобто μ.

Таким чином, основний зміст закону великих чисел полягає в тому, що при збільшенні $n$ окремих вибірок відбувається взаємне погашення індивідуальних відхилень від деякого рівня, характерного для всієї сукупності в цілому. **Закон великих чисел** – один з виразів діалектичного зв'язку між випадковістю і необхідністю.

**Розподіл $X$ малих вибірок.** Коли вибірки є досить великими за обсягом, розподіл їх середніх арифметичних є нормальним. Однак якщо вибірки малі ($n < 30$), то виникають великі неточності щодо можливості судження за такими вибірками про генеральну сукупність. Водночас, у біологічних дослідженнях нерідко доводиться стикатися із вибірковими сукупностями, що складаються з дуже обмеженої кількості варіантів або спостережень. Виникає питання про те, які в цих випадках можуть бути закономірності розподілу вибіркових середніх арифметичних. Відповідь на нього практично дав англійський математик Вільям Сілі Госсет, який працював під псевдонімом Стьюдент, і тому вивчений ним розподіл ймовірностей отримав назву ***t*-розподілу за Стьюдентом.**

Теоретичне обґрунтування закону розподілу відкритого Стьюдентом, блискуче надав математик Рональд Фішер. Важливим є те, що закон може бути використаний і при дуже малих кількостях варіантів.

**Критерій t за Стьюдентом - Фішером** є наступним:

$$t = X - \mu / m_x$$

Виявилося, що розподіл значень *t* відрізняється від нормального, при цьому тим сильніше, чим менше *n*. Тому і ймовірності знаходження вибіркових середніх в межах певних значень *n* значно знижуються порівняно із нормальним розподілом. У практичній роботі треба виходити з певних рівнів значущості, тому були складені робочі таблиці, за якими можна визначати мінімальне значення, що обов'язково потрібне для даної ймовірності (за Петром Фомичем Рокицьким).

**Визначення необхідного обсягу вибіркової сукупності.** На практиці біологічних досліджень часто виникає питання про те, скільки тварин або рослин даного виду потрібно взяти, щоб отримати оптимальне уявлення про популяцію виду за досліджуваною ознакою. Безумовно слід прагнути до більшого числа спостережень, однак очевидно, що чисельність вибірки не може зростати нескінченно. Вона повинна мати якісь раціональні межі, які будуть залежати насамперед від бажаної точності спостереження, тобто допустимої розбіжності між середнім арифметичним (за цією ознакою) вибірки і середнім арифметичним генеральної сукупності, а також від заданої ймовірності і від ступеня однорідності популяції. Бажана точність (позначимо її *Δ,* дельта) – це можливе (за умови прийнятої ймовірності), відхилення *X* від $\mu$, так:

$$\Delta = tm,$$

$$\text{а так як } m = \delta / n,$$

$$\text{то } \Delta = t\,\delta / n.$$

$$\text{Звідси } \mathbf{n = t\,\delta / \Delta}$$

Значення *t* визначається очікуваною ймовірністю результату вибіркового обстеження. Так, при p = 0,997 *t* має дорівнювати 3, а при p = 0,95 можна обмежитися t = 2. Величина Δ береться заздалегідь. Так, наприклад, вивчаючи вагу зайців, можна прийняти, що бажана точність повинна бути в межах 0,2 кг, тобто Δ = 0,2 кг. Дещо важче вирішити питання про величину середнього квадратичного відхилення досліджуваної популяції виду, що є заздалегідь невідомою. У якості її приблизної оцінки можна взяти сигму за даними досліджень, що проводилися раніше або спробувати обчислити її за максимальними і мінімальними значеннями досліджуваної ознаки, маючи на увазі, що варіаційний розмах повинен охоплювати приблизно шість середніх квадратичних відхилень.

**Оцінка достовірності статистичних показників за допомогою середньої помилки.** Оцінка достовірності *X*, середньої або статистичної помилки у статистичному аналізі дуже велика. З одного боку, як було показано вище, вона дозволяє позначити межі для показників генеральної сукупності, наприклад для *y*, а з іншого боку, дає можливість оцінити ступінь

достовірності самих статистичних показників, зокрема середнього арифметичного даної вибіркової сукупності.

Що ж слід розуміти під достовірністю середнього арифметичного? Фактично середнє арифметичне завжди є величиною вибірковою. Тому для судження про його достовірність потрібно порівняти із середнім арифметичним генеральної сукупності. Мірилом достовірності буде нормоване відхилення для обчислення якого можна використовувати наведену вище формулу.

Водночас, виникає питання про те, звідки ж взяти величину *p*? Можливі два випадки. У першому випадку *y* являє собою певну, відмінну від нуля, величину, значення якої можна приблизно припустити за іншими даними. Припустимо, що вивчали жирність молока 10 корів. Було отримано наступні показники: x = 3,7 %; σ = 0,28 %; m = 0,09 %. Якщо при цьому раніше вивчали жирність молока в інших вибірках і отримували різні значення вибіркових середніх, то можна обчислити середнє з цих середніх. Припустимо, що вона буде дорівнювати 4,0 %. Можна прийняти її за $\mu$, тоді $t$ = 3,7-4,0 / 0,09 = 3,3. При малому $n$ (= 8) слід перевірити достовірність за таблицями Рокицького. Так, ймовірність достовірності (p = 0,987) цілком достатня.

Загалом можна сказати, що *x*, обчислені для більшості біологічних показників навіть на порівняно малих за розмірами вибіркових сукупностях, найчастіше будуть досить достовірними, якщо тільки ряд не дуже розтягнутий. Однак може вийти інакше, якщо доводиться оперувати експериментальними даними, в яких фігурують якісь умовні або відносні величини, частина останніх може мати і негативний знак. Тоді встановлення достовірності *x* є абсолютно необхідним.

**Нульова гіпотеза.** Метод середньої помилки дозволяє порівнювати між собою будь-які дві групи тварин або рослин, наприклад: дві вибіркові сукупності, взяті із природної, невивченої популяції; вибірку з якоїсь вже відомої групи і групи, з якої ця вибірка взята; дослідну та контрольну групи при постановці дослідів – і встановити, наскільки достовірні відмінності між їх статистичними показниками (середніми арифметичними, дисперсії та ін.). Загальні принципи порівняння ґрунтуються на аналізі так званої **нульової гіпотези**. Відповідно до цієї гіпотези, спочатку приймається, що між даними показниками (або групами, на основі яких вони отримані) достовірної відмінності немає, тобто обидві групи разом складають один і той же однорідний матеріал або одну сукупність. Статистичний аналіз повинен привести або до відхилення нульової гіпотези, якщо доведена достовірність отриманих відмінностей, або до її збереження, якщо достовірність відмінностей не доведена, тобто відмінності визнані випадковими. Але так як всі статистичні показники і відмінності між ними характеризуються певними рівнями значущості, то відкидання нульової гіпотези має бути пов'язане із прийняттям певного рівня значущості. Так, якщо визнаний необхідним рівень значущості 0,01 і якщо ймовірність достовірності *p* даного статистичного показника або різниці між показниками не задовольняє цій

умові, оскільки вона є нижчою 0,99 (наприклад, 0,97, 0,91, 0,88 тощо), то немає підстав для відкидання нульової гіпотези. Її треба вважати правильною принаймні до тих пір, поки нові дані не дадуть можливості її спростувати, довівши, що існуючі відмінності не є чисто випадковими.

Звичайно, і в тому випадку, коли нульова гіпотеза вважається спростованою, є певний шанс, що вона в дійсності вірна. При рівні значущості 0,01 цей шанс становить 1 на 100, тобто в 1% випадків відкидання нульової гіпотези було помилкою. Якщо досягнуто рівня значущості не 0,01, а 0,001, то впевненість у тому, що нульова гіпотеза дійсно відкинута правильно, різко зростає (лише 1 шанс на 1000 випадків, що вона все ж вірна). При p = 0,05 впевненість у правильності поставленого висновку становить лише 95 випадків зі 100, а в 5 – можливий неправильний результат. Таким чином, якщо отримані дані характеризуються рівнем значущості P <0,05, то немає підстав відхиляти нульову гіпотезу. Якщо P> 0,05 – нульова гіпотеза спростована.

Але значно невизначенішим є стан речей, якщо результати аналізу або порівняння задовольняють рівнем значущості 0,05, але не задовольняють рівнем значущості 0,01. Надійне судження виявляється неможливим. Очевидно, що в таких випадках повинні бути проведені додаткові досліди, щоб вирішити чи слід відкидати нульову гіпотезу, чи ні. Водночас, треба мати на увазі, що збереження нульової гіпотези ще не означає її правильності. Може виявитися все ж, що вона є неправильною, збереження ж нульової гіпотези залишає питання відкритим.

Наведена вище оцінка достовірності середнього арифметичного вибіркової сукупності також є перевіркою нульової гіпотези. Згідно з нульовою гіпотезою, $x = 0$. Необхідно довести, що $x$ достовірно відрізняється від нуля. При достатньому доказі, що задовольняє прийнятому рівні значущості, нульова гіпотеза відкидається, тобто визнається достовірність $x$. Якщо це зробити не вдалося, залишається правильною нульова гіпотеза (тобто недостовірність $x$) і надалі, тобто до нових дослідів.

**Оцінка достовірності різниці між середніми арифметичними двох вибіркових сукупностей.** Якщо отримана різниця між середніми арифметичними двох генеральних сукупностей, то, очевидно, не може стояти питання про статистичної помилку у цій різниці. Ця різниця завжди достовірна, навіть якщо вона і дуже мала. Інша річ, якщо порівнюються дві вибіркові сукупності, наприклад: дві групи морських свинок, які зазнавали впливу хімічних речовин або фізичних чинників, дві групи корів, порівнювані за надоями і взяті з однієї породи, господарства тощо. У цих випадках різниця між середніми має свою статистичну помилку, з якою її можна порівняти і встановити, чи достовірна ця різниця чи ні. Нульова гіпотеза в даному випадку буде зводитися до того, що дві досліджувані вибіркові сукупності походять з однієї і тієї ж генеральної сукупності, і що різниця між їх середніми арифметичними – випадкова, тобто, знаходиться у межах помилки вибірковості.

Щоб мати право відкинути нульову гіпотезу, треба довести, що різниця між середніми арифметичними достовірна, тобто задовольняє необхідному рівню значущості. Для встановлення достовірності різниці між середніми арифметичними треба скористатися нормованим відхиленням. Нормоване відхилення прийме таку форму:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{(\bar{x}_1 - \bar{x}_2)}}.$$

У дійсності ж, формула для t повинна бути дещо складнішою:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{(\bar{x}_1 - \bar{x}_2)}}.$$

Але, в даному випадку потрібно брати за основу нульову гіпотезу про те, що два значення вибіркових середніх арифметичних взяті з однієї генеральної сукупності, тоді $\mu 1 = \mu 2$ і права частина чисельника рівна нулю. Власне, чисельником є різниця між середніми арифметичними двох груп (знак різниці не має значення). Її можна позначити скорочено буквою *d*. У знаменнику ж – середня помилка цієї різниці, тобто $m_{x1} - m_{x2}$ або у скороченому вигляді $s_d$. У такому випадку:

$$t = \frac{d}{s_d}.$$

Водночас, слід розуміти, що існує два способи визначення середньої помилки різниці. Перший з них застосовується, коли обидві порівнювані групи мають досить велике число членів, понад 30 особин у кожній. Середня помилка різниці визначається тоді за формулою:

$$s_d = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2}.$$

*Наприклад,* є необхідність порівняти за кількістю меду 2 пасіки. В одній пасіці n1 = 50. В іншій – n2 = 40. Середня кількість зібраного меду і помилка для першої групи: $X1 \pm mx1$ = 2100 ± 120 кг; для другої групи: $X2 \pm mx2$ = 2635 ± 140 кг. Різниця між середніми зборами меду 2 груп:

$$d = \bar{x}_2 - \bar{x}_1 = 2635 - 2100 = 535 \text{ кг.}$$

Помилка різниці:

$$s_d = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2} = \sqrt{140^2 + 120^2} = 184 \text{ кг.}$$

Таким чином, $d \pm md$ = 535 ± 184 кг, а t = 2,91. Враховуючи дані таблиці нормального інтегралу ймовірності (за Рокицьким) знаходимо, що в цьому випадку ймовірність достовірності дуже велика – 0,9963.

Проте, за відсутності таблиці можна скористатися **правилом трьох сигм:** якщо різниця перевищує свою помилку майже в три рази, вона достовірна з ймовірністю не менше 0,991. Але зі сказаного вище очевидно, що в такому високому значенні t немає потреби. Якщо n> 30, то t = 2,58 гарантує достовірність різниці з ймовірністю 0,99.

При порівнянні двох груп з малими, і, особливо з неоднаковими об'ємами, помилка різниці буде визначатися за дещо іншою формулою:

$$s_d = \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} \left( \frac{n_1 + n_2}{n_1 \cdot n_2} \right)}.$$

Статистичний сенс цієї формули полягає в тому, що не можна користуватися просто готовими середніми помилками, обчисленими заздалегідь для двох порівнюваних груп, а потрібно спочатку скласти суми квадратів відхилень обох груп, тобто отримати об'єднану суму квадратів відхилень, а потім визначити **варіанту об'єднаних рядів** (шляхом ділення об'єднаної суми квадратів на суму чисел ступенів свободи обох груп) і, нарешті, після множення на n1 + n2 / n1 x n2 і вилучення квадратного кореня отримати помилку різниці.

*Приклад.* На двох групах щурів було закладено дослід з порівняння впливу різних раціонів на ступінь зростання. Щури 1 групи (12 од.) отримували раціон з високим вмістом білка, щури другої групи (7 од.) – з низьким. Прирости за 56 днів досліду для кожного щура складали (в г): перша група – 134, 146, 104, 119, 124, 161, 107, 83, 113, 129, 97, 123; друга група – 70, 118, 101, 85, 107, 132, 94.

Після обробки даних за допомогою однієї з формул для сум квадратів отримаємо: d = X1 - X2 = 19 г .; Σ (x - X1) 2 = 5302, Σ (x - X2) 2 = 2575, тоді загальна сума квадратів дорівнює 7584, а ступеня свободи *df* = 17. Застосувавши зазначені вище формули отримаємо *t* = 1,89. За таблицею Рокицького знаходимо, що (при *df* = 17 і рівні значущості 0,05) *t* повинно бути не менше 2,11, а отже отримане значення *t* нижче табличного. Для уточнення ймовірності достовірної різниці скористаємося таблицею інтегралів. З неї видно, що *t* = 1,89 відповідає ймовірності 0,92, тобто рівень значущості 0,08. Таким чином, можна вважати, що різні раціони не призвели до поділу популяції щурів за приростом ваги на дві популяції, що достовірно відрізняються одна від одної, інакше кажучи, нульова гіпотеза не може бути відкинута. Звичайно, причиною такого результату може бути те, що дослідні групи були занадто малими. Можливо якщо їх збільшити, буде отримана достовірніша різниця між групами щурів, які перебували на різних раціонах годівлі.

## *Лекція 5. Основи дисперсійного аналізу.*

**Сутність і метод дисперсійного аналізу.** У попередніх лекціях було розглянуто методи оцінки відмінностей двох вибірок шляхом порівняння їх середніх *μ1* і *μ2* та стандартних відхилень. Проте, у дослідженнях часто доводиться мати справу не з двома, а зі значно більшим числом вибірок і, зазвичай, ці вибірки відносяться до різних сукупностей. Наприклад, це можуть бути групи рослин, обробка яких включала різні добрива або випадок, коли у досліді ставиться мета статистично оцінити ефект заходу. На початку 1950-х років Р. Фішер розробив критерій і метод для такої оцінки, що призвело до значного подальшого розвитку теорії планування експерименту і статистичної оцінки його ефекту.

Статистичний сенс завдання з оцінки ефекту заходу у багатогруповому досліді полягає у перевірці значимості відмінностей у групових середніх оцінюваного показника на основі порівняння дисперсій.

Для розкриття суті методу оцінки ефекту заходу, тобто дисперсійного аналізу, розглянемо спочатку аналіз декількох вибірок, взятих із загальної сукупності. Такий дослід називають **умовним експериментом**.

Дж. У. Снедекор (1961 р.) утворив 4 вибірки (a = 4) із загальної сукупності даних по приросту 511 тварин. Кожна з груп включала n = 5 спостережень (або повторень). Середня для сукупності μ = 30, а дисперсія σ2 = 100. Результати досліду наведені у табл. 5.1.

Таблиця 5.1. **Прирости (у фунтах) 4 груп тварин**

| **Група** | **Приріст *X*** | **Сума ΣX** | **Середнє μ** | **$\Sigma X^2$** | **$\Sigma(X)^2$ / n** | **$\Sigma x^2$** |
|---|---|---|---|---|---|---|
| 1 | 40,24,46,20,35 | 165 | 33 | 5917 | 5445 | 472 |
| 2 | 29,27,20,39,45 | 160 | 32 | 5516 | 5120 | 396 |
| 3 | 11,31,17,37,39 | 135 | 27 | 4261 | 3645 | 616 |
| 4 | 17,21,28,33,21 | 120 | 24 | 3044 | 2880 | 164 |
| **Всього** | | 580 | 29 | 18738 | 16820 | 1918 |

Дані таблиці дозволяють отримати три оцінки дисперсії у сукупності σ2 = 100. Першу оцінку можна отримати на основі усіх 20 спостережень:

$$\sigma = \frac{\sum x^2}{N-1} = \frac{1918}{19} = 100.9, (N = a \cdot n)$$

Друга оцінка виходить із сум квадратів всередині чотирьох груп. Вона відображає варіювання «окремих груп»:

$$\sigma_1 = \frac{\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2}{a \cdot n - n} = \frac{472+396+616+164}{20-4} = 103$$

Групові середні утворюють третю оцінку дисперсії сукупностей. Середній квадрат середніх дорівнюватиме:

$$\frac{(\mu_1-\mu)^2+(\mu_2-\mu)^2+(\mu_3-\mu)^2+(\mu_4-\mu)^2}{n-1}=\frac{(33-29)^2+(32-29)^2+(27-29)^2+(24-29)^2}{4-1}=18$$

Число 18 є оцінкою $\sigma^2$ / 5, тобто оцінкою 20. Кожне середнє представляє 5 спостережень. Отже, третя оцінка $\sigma^2$ буде дорівнювати $\sigma^2$=18*5=90. Вона заснована на 4 групових середніх при n–1=4–1=3 ступенях свободи. Сума квадратів усіх групових середніх буде 90*3=270. Такий результат проведеного загального варіювання на частини та його аналіз називають **дисперсійним аналізом** (табл. 5.2).

Таблиця 5.2. **Дисперсійний аналіз даних щодо приросту тварин**

| **Джерело варіювання** | **Число ступенів свободи** | **Сума квадратів** | **Середній квадрат** |
|---|---|---|---|
| Об’єкти окремих груп | 16 | 1648 | 103 |
| Групове середнє | 3 | 270 | 90 |
| Всього | 19 | 1918 | 100,9 |

1. Сума усіх спостережень:

$$\Sigma X = 40+24+\ldots+21=580.$$

2. Загальна сума квадратів:

$$\sum x^2=\sum X^2-\frac{(\sum X)^2}{N}=40^2+24^2+\ldots+21^2-\frac{580^2}{20}=1918$$

3. Сума квадратів для групових середніх:

$$\frac{\sum(\sum X)^2}{n}-\frac{(\sum X)^2}{a\cdot n}=\frac{165^2+160^2+\ldots+120^2}{5}-\frac{580^2}{20}=17090-16820=270$$

Порівняння середнього квадрата групових середніх (90) і середнього квадрата для об'єктів всередині окремих груп (103) показує незначну їх розбіжність. Перш ніж робити остаточні висновки, доцільно навести схему розрахунків і таблицю аналізу у загальноприйнятому вигляді (табл 5.3).

Таблиця 5.3. **Дисперсійний аналіз даних щодо приросту тварин (загальноприйнята форма)**

| **Джерело варіювання** | **ν** | **$\Sigma x^2$** | **σ** |
|---|---|---|---|
| Загальне | 19 | 1918 | - |
| Групове середнє (факторіальне) | 3 | 270 | 90 |
| Об’єкти окремих груп (випадкове) | 16 | 1648 | 103 |

**Дисперсійний аналіз випадкових вибірок з двох або більшого числа сукупностей.** У більшості визначень дисперсійного аналізу досліджувані варіанти досліду (наприклад, дані дози добрива) впливають на середні. Групи стають вибірками з різних сукупностей. Вважається, що ці сукупності мають різні середні μ, але загальну дисперсію, яка не залежатиме від варіантів досліду. При дисперсійному аналізі середній квадрат для об'єктів оцінює власне $\sigma^2$, як раніше було показано, але середній квадрат групових середніх виявляється перебільшеним в зв'язку з відмінностями між μ. Таблиці 5.4. і 5.5. представляють дані такого експерименту.

Таблиця 5.4. **Висота саджанців тополі, отриманих з живців особин з різними родовими даними (від висоти кожного саджанця віднято 50 см)**

| **Група** | **Висота, см** | | | | | | **Сума** | **Середнє** |
|---|---|---|---|---|---|---|---|---|
| 1 | 64 | 72 | 68 | 77 | 56 | 95 | 432 | 72 |
| 2 | 78 | 91 | 97 | 82 | 85 | 77 | 510 | 85 |
| 3 | 75 | 93 | 78 | 71 | 63 | 76 | 456 | 76 |
| 4 | 55 | 66 | 51 | 64 | 70 | 66 | 372 | 62 |

**Підрахунок:**

1 $\sum X = 64^2 + 78^2 + \ldots + 66^2 = 1770$

2 $\sum x^2 = 64^2 + 78^2 + \ldots + 66^2 - \frac{1770^2}{24} = 3586.5$

**Для середніх:**

$$\frac{432^2 + 510^2 + \ldots + 372^2}{6} - \frac{1770^2}{24} = 1636.5$$

Таблиця 5.5. **Дисперсійний аналіз даних щодо висоті саджанців**

| **Джерело варіювання** | **Число ступенів свободи ν** | **Сума квадратів $\Sigma x^2$** | **Середній квадрат σ** |
|---|---|---|---|
| Загальне | 23 | 3586,5 | - |
| Між групами (факторіальне) | 3 | 1636,5 | 545,5 |
| Варіанти (випадкове) | 20 | 1950 | 97,5 |

**Критерій *F-відношень* дисперсій**. **Висновок про рівність μ.** Отримані дані підводять до питання: чи обумовлюється значна відмінність між середніми квадратами σ1 та σ2 звичайним варіюванням випадкових вибірок з однієї сукупності або воно настільки велике, що слід його приписати впливу вибіркових середніх. Нульова гіпотеза, що відповідає такій постановці питання: $H_0 = \mu 1 = \mu 2 = \ldots = \mu 0$ (середні груп однакові). Для відповіді на ці запитання Р. Фішер запропонував критерій – **відношення дисперсій**, розподіл яких отримано на основі випадкових вибірок з однієї загальної

сукупності. Вище застосування критерію *F* розглядалося для перевірки відмінності в дисперсіях двох малочисельних вибірок.

Дж. Снедекор працював із розподілом, отриманим на основі 100 вибірок по 10 спостережень у кожній, взятих з тієї самої загальної сукупності щодо приросту тварин. Для кожної вибірки за методом, що викладено вище, знайдені *F:*

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

Розподіл 100 значень Р (число ступенів свободи 9 та 90)

| **Інтервал *F*** | 0– | 0,25– | 0,50– | 0,75– | 1,00– | 1,25– |
|---|---|---|---|---|---|---|
| **Число випадків** | 7 | 16 | 16 | 26 | 11 | 8 |
| **Інтервал *F*** | 1,50– | 1,75– | 2,00– | 2,25– | 2,50– | 2,75– |
| **Число випадків** | 5 | 2 | 4 | 2 | 2 | 1 |

Розподіл *F* є несиметричним: 65 значень *F* менше 1. Однак, середнє значення *F*=0.96, тобто близько до очікуваної одиниці. Водночас, 5% значень *F* перевершують 2,25, а 1% вище 2,75. Такою таблицею розподілу Р можна користуватися на практиці. Можна, наприклад, сказати, що при вибірках в 10 одиниць, значення *F* > 2,75 може трапитися внаслідок випадкових причин 1 раз на 100 випадків.

На основі досліджень Р. Фішера отримано теоретичний розподіл *F-критерію* для різних рівнів значимості і для різного числа ступенів свободи. Так, за числа ступенів свободи ν = 3 і ν = 20 маємо 5% -вий рівень критерію F = 3,10. Отриманий у досліді з саджанцями критерій відношення дисперсій:

$$F = \frac{\sigma_1^2}{\sigma_2^2} = \frac{545.5}{97.5} = 5.6 > F_{0.05}.$$

Воно навіть перевищує $F_{0,01}$ =4,9.

На підставі співставлення *F*, отриманого у досліді з табличними значеннями можна сказати, що внаслідок випадкових причин з однієї спільної сукупності ми маємо менше однієї можливості зі 100 отримати вибірку, що дає значення *F* більше, ніж за умов спостережень. Очевидно, що дані аналізованої вибірки належать до сукупності з різними μ. Отже, потрібно дати ствердну відповідь на поставлене вище запитання щодо впливу материнських спадкових якостей на зростання нового покоління, і тоді нульова гіпотеза $H_0$*: μ1 = μ2 = ... = μ0* відкидається.

Такий висновок отримано на основі встановленого, значимо вищого варіювання між груповими середніми, вимірюваного $\sigma^1_2$, порівняно з варіюванням висоти рослин усередині груп, що вимірюється $\sigma^2_2$.

**Дисперсійний аналіз з класифікацією за двома ознаками.** У розглянутому вище прикладі з висотою саджанців була використана

класифікація тільки за однією ознакою. Проте, часто, дисперсійний аналіз застосовується і у випадках, коли об’єкт класифікується за кількома ознаками. Наприклад, є угруповання за двома ознаками (факторами), власне, значимість яких і перевіряють. Наявні такі результати спостережень X щодо впливу добрив (B1 і B2) на ґрунтах із різним якісним складом (A1 і A2) (табл. 5.6.).

Таблиця 5.6. **Результати спостережень X**

| **Добриво** | **Ґрунт** | | | |
|---|---|---|---|---|
| | **A1** | **A2** | | |
| **B1** | 8, 12<br>$\mu_{11}=10;\dots\sum$ | 1, 3<br>$x_{11}^2=8$ $\mu_{21}=2;\dots\sum x$ | $_{21}^{2}=2$ $\mu_{B1}=\frac{24}{4}=6$ | $\sum x_B^2=0$ |
| **B2** | 3, 4, 5<br>$\mu_{12}=4;\dots\sum x_{12}^2=2$ | 6, 8, 10<br>$\mu_{22}=8;\dots\sum x_{12}^2$ | $\mu_{B2}=\frac{36}{6}=6$ | |
| **Уся група** | $\mu_{A1}=\frac{32}{5}=6.4$ | $\mu_{A2}=\frac{28}{5}=5.6$ | $\mu=6;\dots\sum x^2=108$ | |
| | $\sum x_A^2=1.6$ | | | |

- Числа 8, 12, 3, 4, 5, 1, 3, 6, 8, 10 – значення результативної ознаки – X.
- *μ11, μ12, μ21, ... μ22* – часткові середні в клітинах; вони отримані за формулою:

$$\mu=\frac{\sum X_i}{n}$$

- *μA1, μA2* – середні для 1 і 2-ї груп ґрунтів;
- *μB1, μB2* – аналогічно для відповідних груп добрив.
- $\sum x_{11}^2 \cdots \sum x_{22}^2$ – суми квадратів відхилень варіантів від середніх у клітинах.

У дослідах, подібних до розглянутого вище, зазвичай формують перевірку наступних гіпотез:

1. Чи відрізняються значимо за своїм ефектом на ріст рослин ґрунти A1 і A2?

Відповідь на це питання містять середні для двох груп ґрунтів: μA1 = 6,4 і μA2 = 5,6. Відмінності такого роду, що пов'язані з невід'ємними якісними чинниками середовища, в літературі про дисперсійний аналіз називаються *ефектом середовища*.

2. Наскільки значно різниться ефект двох добрив B1 і B2?

У даному випадку, відповідь міститься у кінцевих двох рядків: μB1 = μB2 = 6,0. Відмінності, пов'язані з процесом виробництва, в даному випадку з добривом, називають *ефектом обробки*.

3. Чи впливають добрива на ріст рослин в однаковій мірі на обох ґрунтах?

Відповідь на третє питання слід шукати у середніх по клітинам μ11, μ12, μ21, ... μ22. Видно, що використання добрива B1 на ґрунті A1 призвело до середньої μ11 = 10, тоді як на ґрунті A2 середня μ21 = 2. Добриво B2

характеризується зворотним результатом: μ12 = 4; μ22 = 8. Відповідь на третє питання можна охарактеризувати як ефект взаємодії чинників АВ.

Для підтвердження або спростування відповідей на попередні 3 питання доцільно висунути 3 нульові гіпотези:

- гіпотеза H*a*- середні стовпчиків не відрізняються один від одного;
- гіпотеза H*b*- середні рядків не відрізняються один від одного;
- гіпотеза H*ab*- взаємодія *ab* відсутня.

**Компоненти загальної суми квадратів.** Загальна сума квадратів:

$$\sum x = \sum\left(8^2 + 10^2 + \ldots + 6^2 + 8^2 + 10^2\right) - \frac{60^2}{10} = 468 - 360 = 108.$$

Цю суму квадратів поділяємо на компоненти, що вимірюють вплив двох випробовуваних чинників, їх взаємодію, а також вплив масштабного числа випадкових чинників або «компонента помилки» – міри коливань внаслідок гри випадку. Сума квадратів, що відповідає кожному з принципів класифікації, обчислюється так само, як і при однофакторному комплексі – як сума квадратів відхилень кожної групової середньої від загальної середньої (з урахуванням ваги *ni* кожної середньої):

- для чинника ґрунту – $\sum x_A^2 = (6.4-6)^2 \cdot 5 + (5.6-6)^2 \cdot 5 = 1.6$
- для чинника добрива – $\sum x_B^2 = (6-6)^2 \cdot 5 + (6-6)^2 \cdot 5 = 0.$

**«Компонент помилки»** – незалежний від двох покладених в основу класифікації, принцип, являє собою суму квадратів всередині всіх чотирьох елементів:

$$\sum\sum x^2 = 8+2+2)8 = 20$$

Ця сума квадратів, розділена на відповідне число ступенів свободи, приймається в якості запобіжного впливу від випадкових чинників.

Сума трьох компонентів: $\sum x_A^2 + \sum x_B^2 + \sum\sum x^2 = 1.6 + 0 + 20 = 21.6.$

Віднімаючи цей результат від загальної суми $\sum x^2 = 108,$ отримаємо залишок рівний 86,4. Цей залишок можна визначити як «**залишкову міжгрупову мінливість**». Він буде вимірювати взаємодію АВ.

Для ступенів свободи знайдених 4-х компонентів, отримуємо наступні залежності (a – число стовпців, b – число рядків).

| | Ступінь свободи: |
|---|---|
| Між рядками | b-1 |
| Між стовпчиками | a-1 |
| Для взаємодії | (a-1) (b-1) |
| Всередині клітин | N-ab |
| Для кінцевого результату | N-1 |

Дані аналізу підтверджують нульову гіпотезу Ha = 0; Hb = 0, але не узгоджуються з нульовою гіпотезою Hab = 0. Ця гіпотеза відкидається на 1% -му рівні значущості, тобто при ймовірності p = 0,99 (табл. 5.7.).

Таблиця 5.7. **Дисперсійний аналіз**

| **Джерело варіювання** | **Ступінь свободи ν** | **Сума квадратів $\Sigma x^2$** | **Середній квадрат (дисперсія) σ** | **F** |
|---|---|---|---|---|
| Фактор А (ґрунт) | 1 | 1,6 | 1,6 | $0,5<F_{0,05}=6,0$ |
| Фактор В (добрива) | 1 | 0 | - | |
| Взаємодія АВ | 1 | 86,4 | 86,4 | $26,2<F_{0,05}<F_{0,01}=13,4$ |
| Всередині клітин (помилка) | 6 | 20 | 3,3 | |
| Усього | 9 | 108 | | |

Отже, згідно з даними результатами аналізу можна зробити висновок, що 2 види добрив В1 і В2 тісно взаємодіють з ґрунтом, тобто проявляють ефект залежно від типу ґрунту. Проте, коректніше говорити так – ґрунти А1 і А2 по-різному реагують на добрива.

## *Лекція 6. Кореляційний аналіз*

**Поняття про кореляцію.** Викладені у попередніх розділах методи аналізу дають можливість вивчати варіацію тварин чи рослин за кожною окремою ознакою – вазі/масі, замірам/висоті, плодючості/продуктивності тощо. Однак, у ряді випадків важливо розуміти, якою саме є залежність між варіацією двох або навіть кількох ознак; чи змінюються дві ознаки самостійно або незалежно одна від одної, чи може бути, що варіація однієї ознаки певною мірою пов'язана із варіацією іншої.

Існують дві категорії зв'язків, або залежностей між ознаками: **функціональні і кореляційні, або статистичні.** При функціональних залежностях кожному значенню однієї змінної величини відповідає одне цілком певне значення іншої змінної. Такі залежності спостерігаються в математиці і фізиці, оскільки різні вимірювальні прилади засновані на функціональних залежностях. Так, висота ртутного стовпчика в термометрі дає точну і однозначну відповідь щодо температури повітря або води. Між радіусом кола К і його довжиною С існує функціональна залежність за відомою з елементарної геометрії формулою $C = 2\pi R$. Інакше кажучи, кожному значенню Х відповідає строго певне значення Y. Аналогічно і напруження нитки в електричній лампочці визначається напругою.

Водночас, поряд із функціональними існують статистичні зв'язки, при яких чисельному значенню однієї змінної відповідає багато значень іншої змінної. Наприклад, між кількістю внесених на поле добрив і врожайністю пшениці існує безперечна залежність. Проте, це не означає, що певній кількості добрив відповідає строго певна величина врожаю. У формуванні врожаю на цій ділянці поля приймає участь багато чинників (склад і структура ґрунту, способи внесення добрив, глибина їх закладення, відмінності у методах посіву тощо). У багатьох дослідженнях потрібно вивчити кілька ознак у їх взаємних зв'язках. Якщо проводити таке дослідження відносно двох ознак, то можна помітити, що мінливість однієї ознаки знаходиться в деякій відповідності до мінливості іншої.

У деяких випадках така залежність проявляється настільки сильно, що при зміні першої ознаки на певну величину, завжди змінюється і друга ознака на певну величину, тому кожному значенню першої ознаки завжди відповідає цілком певне, єдине значення другої ознаки. Такі зв'язки називаються **функціональними**. Зустрічаються функціональні зв'язки у фізичних і математичних узагальненнях. Площа трикутника точно визначається його висотою і основою, довжина кола – радіусом, швидкість падіння є функцією часу падіння і прискорення сили тяжіння, швидкість протікання певної хімічної реакції залежить від температури і т.д. Проте, необхідно врахувати, що функціональні зв'язки зустрічаються тільки в ідеальних умовах, коли передбачається, що ніяких сторонніх впливів немає.

При вивченні живих об'єктів – диких і культурних рослин, тварин, мікроорганізмів – доводиться мати справу зі зв'язками іншого роду. Живий

організм розвивається у різнобічних зв'язках з умовами його життя, під дією нескінченно великого числа чинників, які по-різному визначають розвиток різних ознак. У живих об'єктів зв'язок між будь-якими двома ознаками настільки часто і сильно порушується або модифікується, що не завжди можуть бути легко виявлений. У рослин, тварин і мікроорганізмів зв'язок між ознаками зазвичай проявляється особливим чином: кожному значенню першої ознаки відповідає не одне значення другої ознаки, а цілий розподіл цих значень за умови цілком певних основних показників цього часткового розподілу – середньої величини і ступеня різноманітності. Такий зв'язок називається **кореляційним зв'язком або просто кореляцією.**

Кореляційний зв'язок, наприклад, між вагою тварин та їх довжиною виражається в тому, що кожному значенню довжини відповідає певний розподіл ваги (а не одне значення ваги), і зі збільшенням довжини збільшується і середня вага тварин. Кореляційний зв'язок не є точною залежністю однієї ознаки від іншої, тому він може мати різну ступінь – від повної незалежності до дуже сильного зв'язку. Крім того, характер зв'язку між різними ознаками може бути різний. Тому виникла необхідність визначати форму, напрямок і ступінь кореляційних зв'язків.

За формою кореляція може бути прямолінійною і криволінійною, у напрямку – прямий і зворотній. Ступінь кореляції вимірюється різними показниками, введеними для встановлення сили зв'язку між кількісними і якісними ознаками. Такими показниками є коефіцієнт кореляції $r$ та кореляційне відношення $\eta$.

Зобразити кореляційний зв'язок двох ознак можна трьома способами:

1. За допомогою кореляційного ряду, що складається з ряду пар значень, у тому числі одне значення відноситься до першої ознаки, а інше в цій парі – відноситься до другої, що пов'язана з першим. На рис. 1 показані схеми кореляційних рядів при п'яти ступенях кореляційної зв'язку.
2. За допомогою кореляційної решітки, у якій кожній особині відповідає певна клітинка. На рис. 1 показана схема кореляційних решіток для п'яти ступенів кореляційного зв'язку між двома ознаками. Значення першої ознаки нанесені по осі абсцис, значення другої – на осі ординат.
3. За допомогою лінії регресії, абсциси якої пропорційні значенням першої ознаки, а ординати – значенням другої ознаки, що кореляційно пов'язані із першим. На рис. 6.1 показані схеми ліній регресії для п'яти ступенів кореляційного зв'язку між двома ознаками.

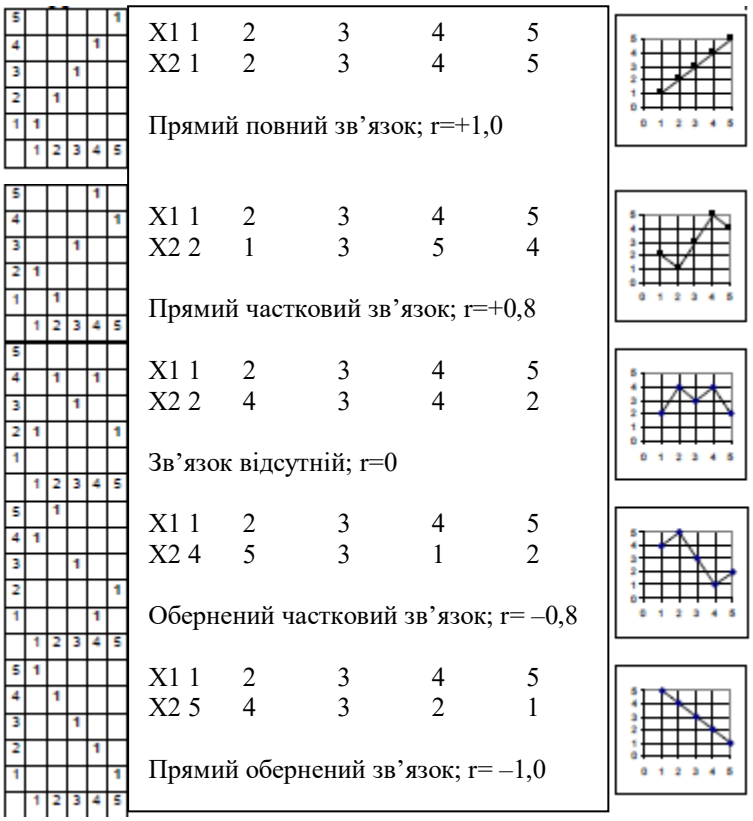



**Рис. 6.1** Схеми кореляційних рядів при п'яти ступенях кореляційного зв'язку.

**Коефіцієнт кореляції.** Коефіцієнт кореляції вимірює ступінь і визначає напрямок прямолінійних зв'язків. Лінійна залежність між ознаками – це такий зв'язок, при якому рівномірним змінам першої ознаки відповідають рівномірні (у середньому) зміни другої ознаки при незначних і хаотичних відхиленнях від цієї рівномірності. Наприклад, при збільшенні довжини тіла на кожен сантиметр ширина збільшується у середньому на 0,7 см. При графічному зображенні прямолінійних зв'язків (рис. 1) (якщо по осі абсцис відкласти значення першої ознаки, по осі ординат – другої, а отримані точки з'єднати) виходить пряма або така крива, середнє значення якої проходить по прямій. При зображенні прямолінійних кореляційних зв'язків у формі кореляційних решіток (рис. 1) частоти всередині розташовуються у формі еліпса. Велика вісь цього еліпса проходить або по діагоналі від кута найменших значень (при позитивному кореляційному зв'язку), або по діагоналі від кута, де сходяться найменші значення однієї ознаки і найбільші значення іншої, до протилежного кута (при негативному кореляційному зв'язку).

При вимірі ступеня зв'язку між різними ознаками часто доводиться порівнювати величини, виражені в різних одиницях виміру. Наприклад, при вимірюванні зв'язку між вагою тварини та його довжиною треба зіставити кілограми ваги з сантиметрами довжини. В інших випадках зміни об’єму зіставляються зі змінами віку, зміни ваги руна в кілограмах зі змінами вмісту в ньому жиропоту у відсотках, довжина ніг у сантиметрах зі швидкістю бігу в хвилинах і т. д.

Проводити такі порівняння виявилося можливим шляхом використання нормованого відхилення, що обчислюється за формулою:

$$\bar{x}_i = \frac{X_i - \mu}{\sigma}.$$

Нормоване відхилення служить універсальною і неіменованою мірою розвитку ознак. Ці властивості нормованого відхилення і дозволили сконструювати основний показник кореляційної зв'язку – коефіцієнт кореляції. Основна формула, яка розкриває сутність цього показника, має зовсім просту структуру:

$$r = \frac{\sum \bar{x}_1 \cdot \bar{x}_2}{\nu}$$

, де

r – коефіцієнт кореляції;

$\dot{x}_1 \bullet \dot{x}_2$ – нормовані відхилення дат за першою та другою ознакою;

ν – число ступенів свободи, що в даному випадку дорівнює числу порівнюваних пар мінус одна.

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

- Якщо обидві ознаки змінюються паралельно, то сума їх нормованих відхилень дає позитивну величину. Якщо при збільшенні однієї ознаки інший зменшується, то доводиться множити позитивні числа на негативні і вся сума утворених нормованих відхилень дає негативну величину. Тому коефіцієнт кореляції може визначати напрямок зв'язку: при прямих зв'язках він позитивний, а при зворотних зв'язках – негативний.
- При повних зв'язках, коли зміни обох ознак строго відповідають один одному і кореляційний зв'язок перетворюється у функціональний, сума утворених нормованих відхилень стає рівною числу ступенів свободи:

  $$\sum \bar{x}_1 \cdot \bar{x}_2 = \nu = n - 1$$

  Тому максимальне значення коефіцієнта кореляції дорівнює 1; для позитивних або прямих зв'язків:

  $$r_{max} = \frac{\sum \bar{x}_1 \cdot \bar{x}_2}{\nu} = \frac{+\nu}{\nu} = +1.0$$

  Для негативних або обернених зв'язків:

  $$r_{min} = \frac{\sum \bar{x}_1 \cdot \bar{x}_2}{\nu} = \frac{-\nu}{\nu} = -1.0$$

- При повній відсутності кореляційного зв'язку між ознаками сума утворених нормованих відхилень дорівнює нулю, і тому коефіцієнт кореляції у цих випадках теж дорівнює нулю:

  $$r_{min} = \frac{\sum x_1 \cdot x_2}{\nu} = \frac{0}{\nu} = 0$$

Граничні значення коефіцієнта кореляції (r = + 1,0; r = 0,0; r = -1,0) на практиці зустрічаються вкрай рідко. П'ять основних видів прямолінійного кореляційної зв'язку, відповідні коефіцієнтам кореляції +1,0; +0,8; 0,0; -0,8 і -1,0, показані на рис. 1. Основна формула коефіцієнта кореляції чітко

розкриває сутність цього показника, але для роботи є дуже незручною, особливо при багаточисельних групах. Тому, на сьогодні, розроблено різноманітні робочі формули для практичних розрахунків у різних умовах - для малих і великих груп при малозначних і багатозначних варіантах. Всі ці формули дають однаковий результат і застосування будь-якої із них обумовлюється тільки зручністю і простотою необхідних обчислень. Найчастіше у біологічних дослідженнях використовують дві формули, запропоновані для малих груп:

$$r = \frac{\sum X_1 \cdot X_2 - \frac{\sum X_1 \sum X_2}{n}}{\sigma_1 \cdot \sigma_2}$$

де X1, X2 - дати першої і другої ознак; N - число порівнюваних пар дат, або число об'єктів, у яких виміряно по дві ознаки; σ1, σ 2 - стандартні відхилення за першою та другою ознакою.

Застосовується коефіцієнт кореляції в тих випадках, коли необхідно дізнатися напрямок і силу зв'язку між ознаками, причому заздалегідь відомо, що цей зв'язок може вважатися прямолінійним, або коли потрібно з'ясувати ступінь саме прямолінійного зв'язку. При цьому краще проводити два етапи дослідження: 1) розгляд кореляційної решітки; 2) розрахунок коефіцієнта кореляції або по цій же решітці, або безпосередньо по датам. Уже власне вигляд кореляційної решітки дозволяє приблизно встановити напрям і ступінь прямолінійних зв'язків, а також характер криволінійних зв'язків. У разі відомого досліду за видом кореляційної решітки можна отримати перше уявлення про особливості і силу зв'язку між досліджуваними ознаками. Полегшує вирішення цього завдання схема ступенів прямолінійної кореляції, показана у табл. 6.1. У цій схемі наведено стандартні кореляційні розподіли 50 особин при різних ступенях прямолінійного зв'язку з дев'яти градацій від r = + 1,0 до r = -1,0.

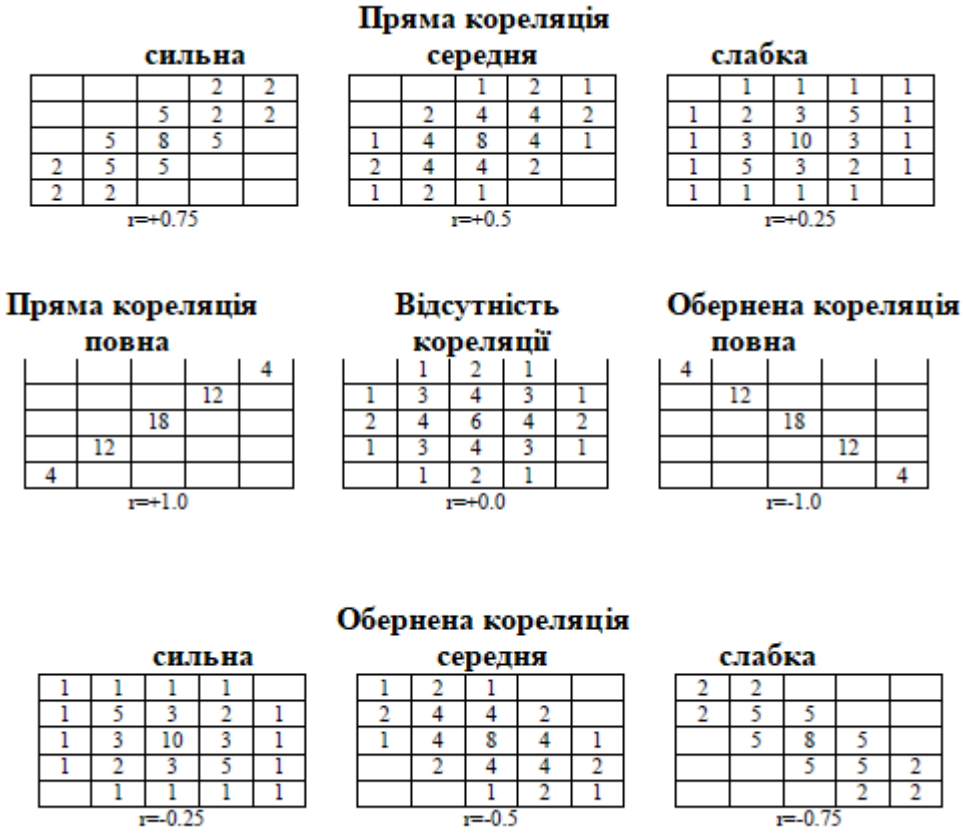
Схемою ступенів прямолінійної кореляції можна користуватися як еталоном для початкового орієнтовного віднесення досліджуваного зв'язку до однієї з умовних ступенів («сильна», «середня», «слабка») тільки згідно з формою кореляційної решітки. У деяких випадках така груба оцінка буває достатньою для з'ясування попередніх питань дослідження.

**Помилка коефіцієнта кореляції.** Як і будь-яка вибіркова величина, коефіцієнт кореляції має свою помилку репрезентативності, яка обчислюється для великих вибірок за формулою:

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

r - коефіцієнт кореляції у генеральній сукупності, з якої взята вибірка;
n - чисельність вибірки, тобто число пар значень, за якими обчислювався вибірковий коефіцієнт кореляції.

Таблиця 6.1. **Схема ступенів прямолінійної кореляції**




Оскільки у чисельнику формули помилки вибіркового коефіцієнта кореляції стоїть квадрат генерального коефіцієнта кореляції, то ця формула може застосовуватися лише у виняткових випадках, коли заздалегідь відома або передбачається ступінь кореляції у генеральній сукупності. *Наприклад.* Для перевірки гіпотези про те, що коефіцієнт кореляції між дітьми і батьками r = + 0,5, була порівняна плодючість 226 лисиць і їх дочок у відповідному віці і в подібних умовах. Коефіцієнт кореляції виявився рівним 0,45. Підтверджує або спростовує цей результат дану гіпотезу?

У даному випадку різниця між вибірковим і генеральним коефіцієнтами d = + 0,45 - (+ 0,50) = - 0,05, а її помилка дорівнює помилці вибіркового коефіцієнта, так як генеральні величини не мають помилок репрезентативності. Для обчислення помилки коефіцієнта кореляції є можливість застосувати точну формулу з генеральним коефіцієнтом у чисельнику:

$$s_r = \frac{1-0.5^2}{\sqrt{225}} = \frac{0.75}{15} = 0.05$$

Виявилося, що критерій достовірності різниці $t_{(r-r')} = \frac{0.05}{0.05} = 1$ не перевищує навіть першого порога достовірності ($t_1$ = 2,0, $\beta_1$ = 0,95). Гіпотеза в даному дослідженні не спростована, так як емпіричний коефіцієнт кореляції недостовірно відрізняється від гіпотетичного. У більшості досліджень значення коефіцієнта кореляції у генеральної сукупності невідомо, тому

замість точного значення помилки коефіцієнта кореляції беруть наближене значення:

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

де r - вибіркове значення коефіцієнта кореляції,
n - число порівнюваних пар даних або число об'єктів, у яких виміряні дві ознаки.

Помилка коефіцієнта кореляції використовується для визначення: 1) достовірності вибіркового коефіцієнта кореляції; 2) довірчих меж генерального коефіцієнта кореляції; 3) достовірності різниці двох вибіркових коефіцієнтів кореляції; 4) достовірності різниці між вибірковим і генеральним коефіцієнтом кореляції.

**Достовірність вибіркового коефіцієнта кореляції.** Критерій вибіркового коефіцієнта кореляції визначається за формулою:

$$t_r = \frac{r}{s_r} \geq t_{st}\{\nu = n-2\}$$

де $t_r$ – критерій достовірності коефіцієнта кореляції;
r – вибірковий коефіцієнт кореляції;
n – число пар дат, що корелюються;
$t_{st}$ – стандартне значення критерію Стьюдента, що вираховується згідно із таблицею встановленого числа ступенів свободи і порога ймовірності безпомилкових прогнозів.

При $t \geq t_{st}$ коефіцієнт кореляції є достовірним. В такому випадку з певною ймовірністю можна вважати, що між ознаками, що корелюються є зв'язок і в генеральній сукупності, такий же за знаком, як і у вибірці (прямий або зворотний). При $t < t_{st}$ вибірковий коефіцієнт кореляції недостовірний, що не дає можливості зробити будь-який висновок про зв'язок ознак у генеральної сукупності. Для з'ясування цього питання потрібно провести повторні дослідження на чисельнішому матеріалі. *Наприклад.* При перевірці гіпотези про зв'язок крупноплідності із жирномолочністю був розрахований коефіцієнт кореляції між відсотком жиру в молоці у 50 корів і вагою при народженні телят від цих же корів. У результаті отримано:

- Коефіцієнт кореляції r=+0,21
- Помилка коеф. кор. $s_r = \sqrt{\frac{1-0.21^2}{50-2}} = 0.14;$
- Критерій достовірності $t_r = \frac{0.21}{0.14} = 1.5$; $\nu$=48; $t_{st}$= {2,0-2,7-3,5}.

Вибірковий коефіцієнт виявився фактично недостовірним. На основі проведеного дослідженням не можна очікувати зв'язок між крупноплідністю і жирномолочністю у всіх корів взагалі. Визначення достовірності коефіцієнта кореляції можна значно спростити, використовуючи властивості особливої функції запропонованої Фішером:

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

За допомогою цієї функції можна заздалегідь визначити, при якому обсязі вибірки коефіцієнт кореляції певної величини буде достовірним за необхідного порогу ймовірності безпомилкових прогнозів, за наступною формулою:

$$\hat{n} = \frac{t^2}{z^2} + 3$$

де n - кількість пар значень, достатня для достовірності вибіркового коефіцієнта кореляції,

t - критерій Стьюдента для кожного з трьох порогів ймовірності безпомилкових прогнозів ($\beta_1 = 0{,}95$, $\beta_2 = 0{,}99$, $\beta_3 = 0{,}999$), для великих груп: $t_1 = 1{,}96$, $t_2 = 2{,}58$, $t_3 = 3{,}30$.

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

За цією формулою було розраховане значення z і кількість пар значень, що достатня для достовірності вибіркового коефіцієнта кореляції для кожного із трьох порогів ймовірності безпомилкових прогнозів. У прикладі в вибіркою обсягом n = 50 отриманий коефіцієнт кореляції r = +0,21. При r = 0,21, розраховані три числа: 87-149-242. Це означає що вибірковий коефіцієнт кореляції рівний r = 0,21, може стати достовірним в тому випадку, якщо обсяг вибірки (число корельованих пар даних) буде: для першого порога ймовірності 87, для другого – 149, для третього – 242. Так як фактичний обсяг вибірки n = 50, що далеко не досягає першого максимального порогу, то отриманий коефіцієнт кореляції і виявився недостовірним, що було знайдено і звичайним способом.

**Довірчі границі коефіцієнта кореляції.** Довірчі границі генерального значення коефіцієнта кореляції знаходяться загальним способом за формулою:

$$\bar{r} = r \pm \Delta,$$

де $\bar{r}$ і r – генеральне та вибіркове значення коефіцієнта кореляції;

$\Delta = t*Sr$ – можлива похибка при визначенні генерального параметра;

$t_{st}$ – критерій Стьюдента при числі ступенів свободи $\nu = n-2$;

Sr – помилка коефіцієнта кореляції.

*Наприклад.* При розробці способів визначення ваги устриць певного виду за їх довжиною було виміряно і зважено 200 екземплярів і визначений коефіцієнт кореляції між вагою і довжиною r = + 0,85. Помилка цього коефіцієнта:

$$s_r = \sqrt{\frac{1-0.85^2}{200-2}} = 0.037$$

Число ступенів свободи і критерій Стьюдента:

$\nu = n- 2 = 198$, $t_{st} = \{2{,}0 - 2{,}6 - 3{,}3\}$.

Можлива похибка при прогнозі генерального параметра:
Δ = t*sr = 2,0*0,037 = 0,074.
Довірчі границі:
r̄= + 0.85±0.074 [–не більше +0.85 + 0.074 = 0.92; –не менш 0.85-0.074 =0.78]

Навіть мінімальна межа (гарантований мінімум) виявилася досить високою. Це вказує на можливість практичного використання розкритої закономірності шляхом розробки формули регресії для визначення ваги устриць за їх довжиною з практично високою точністю.

**Достовірність різниці двох коефіцієнтів кореляції.** Достовірність різниці коефіцієнтів кореляції визначається так само, як і достовірність різниці середніх, за звичайною формулою:

$$t_d = \frac{d}{s_d} \geq t_{st}\{\nu = n_1 + n_2 - 4\},$$

де $t_d$ – критерій достовірності різниці коефіцієнтів кореляції;
d = r1-r2 – різниця коефіцієнтів кореляції;

$s_d = \sqrt{s_1^2 \cdot s_2^2}$ – помилки різниці, яка дорівнює кореню квадратному із суми квадратів помилок обох порівнюваних коефіцієнтів кореляції; $s^2 = \frac{1-r^2}{n-2}$

$t_{st}$ – стандартні значення критерію Стьюдента;
ν – число ступенів свободи для різниці коефіцієнтів кореляції, яка дорівнює загальній кількості чисел ступенів свободи обох коефіцієнтів: ν = n1-2 + n2-2 = n1 + n2-4.

*Наприклад.* При розробці способів визначення висоти дерева за його обхватом (по висоті грудей особи, що вимірює) отримані коефіцієнти кореляції між цими ознаками для двох порід дерев:

$$n_1 = 200,\ r_1 = 0{,}60,\ s_1^2 = \frac{1-0.6^2}{198} = 0.0032;$$

$$n_2 = 150,\ r_2 = 0{,}80,\ s_2^2 = \frac{1-0.8^2}{148} = 0.0024.$$

Для з'ясування можливості застосування єдиної формули перерахунку обхвату на висоту треба було з'ясувати: чи достовірна відмінність зв'язку висоти з обхватом між двома досліджуваними породами дерев. Отже, отримано наступні результати:

$$d = 0{,}80 - 0{,}60 = 0{,}20;$$

$$s_d^2 = 0{,}0032 + 0{,}0024 = 0{,}0056,\ s_d = \sqrt{0{,}0056} = 0{,}075;$$

$$t_d = \frac{0.200}{0.075} = 2.7,\ \nu = 200 + 150 - 4 = 346,\ t_{st} = \{2{,}0 - 2{,}6 - 3.3\}.$$

Виявилося, що порівнювані породи досить достовірно (по другому порогу ймовірності) розрізняються за ступенем зв'язку між висотою і обхватом дерева. Тому для цих порід можна користуватися єдиною формулою перерахунку обхвату на висоту.

## *Лекція 7. Регресійний аналіз*

**Різноманіття методів вивчення зв'язків.** Різні залежності є широко поширеними не лише в органічній, а й у неорганічній природі. Їхнє вивчення почали проводити уже давно і на сьогодні вже розроблено значну кількість методів їхньої математичної характеристики. Перший метод це **метод кореляції**.

**Коефіцієнт кореляції** вказує лише на ступінь зв'язку у варіації двох змінних величин або, як іноді кажуть, на поняття тісноти цього зв'язку, але не дає можливості судити про те, як кількісно змінюється одна величина згідно зі зміною іншої. Проте, на це питання дає можливість відповісти інший метод визначення зв'язку між ознаками, що носить назву **метод регресії.**

У сучасній статистиці, зокрема біологічній, коефіцієнтами кореляції користуються рідше, ніж раніше, а от метод регресії набуває дедалі більшого значення. Аналіз взаємовідносин двох змінних величин за допомогою методу регресії часто може дати дуже цінні результати, особливо на практиці. У деяких випадках для висвітлення різних сторін питання потрібно застосовувати і кореляційний, і регресійний методи аналізу.

За умови **простої кореляції** вивчається залежність між мінливістю двох ознак $x$ та $y$. За допомогою регресії ставиться додаткове завдання: встановити як кількісно змінюється одна величина за зміни іншої на одиницю. Оскільки змінних величин дві, то регресія, очевидно, може бути **двосторонньою**:

1. визначення зміни $y$ за зміною $x$;
2. визначення зміни $x$ за зміною $y$.

Саме у цьому полягає основна відмінність методу регресії від методу кореляції. Регресія може бути виражена декількома способами:

- шляхом побудови так званих **емпіричних ліній регресії**,
- шляхом складання рівнянь регресії та побудови **теоретичних ліній регресії**,
- за допомогою обчислення **коефіцієнта регресії**.

Перші два способи дозволяють виразити регресію графічно. Для побудови **емпіричних ліній регресії** можна скористатися звичайними кореляційними решітками, але тут слід замінити межі класів середніми значеннями класів.

**Коефіцієнт прямолінійної регресії.**

**Прямолінійна кореляція** відрізняється тим, що за такої форми зв'язку, кожній з однакових змін першої ознаки відповідає цілком певна і теж однакова в середньому зміна іншої ознаки, пов'язаної з першою або залежною від першої. А, та величина, на яку в середньому змінюється друга ознака, при зміні першої на одиницю виміру, називається **коефіцієнтом регресії**. Розраховується він за такою формулою:

$$R_{2/1} = \frac{\sigma_2}{\sigma_1} \cdot r_{12},$$

де, $R_{2/1}$ – коефіцієнт регресії другої ознаки за першою;
$\sigma_2$ − середнє квадратичне відхилення другої ознаки, яке змінюється у зв'язку із зміною першої;
$\sigma_1$ − середнє квадратичне відхилення першої ознаки, у зв'язку із зміною якого змінюється друга ознака;
$r_{12}$ – коефіцієнт кореляції між першою та другою ознаками.

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

$$s_R = \frac{\sigma_2}{\sigma_1} \cdot s_r = \frac{\sigma_2}{\sigma_1} \cdot \sqrt{\frac{1-r^2}{n-2}}.$$

**Критерій достовірності коефіцієнта регресії** дорівнює критерію достовірності коефіцієнта кореляції:

$$t_R = \frac{R}{s_R} = \frac{\frac{\sigma_2}{\sigma_1} \cdot r_{12}}{\frac{\sigma_2}{\sigma_1} \cdot s_r} = \frac{r}{s_r} = t_r.$$

*Приклад.* Для розробки способу визначення ваги коней без зважування за обхватом грудей було зважено 1618 коней і у кожної особини виміряли обхват. Отримано такі показники: $x$ – обхват грудей, $n$ = 1618, $\mu_x$ = 174 см, $\sigma_x$ = 7,9 см;
$y$− вага, $n$ = 1618, $\mu_y$ = 424 кг, $\sigma_y$ = 56,8 кг.
Коефіцієнт кореляції $r_{x/y}$ = +0,89±0,011.
Коефіцієнт регресії ваги за обхватом дорівнює:

$$R_{y/x} = \frac{\sigma_y}{\sigma_x} \cdot r_{y/x} = \frac{56.8}{7.9}(+0.89) = +6.4$$

Помилка коефіцієнта регресії ваги коней за обхватом грудей дорівнює:

$$s_R = \frac{\sigma_y}{\sigma_x} \cdot s_r = \frac{56.8}{7.9} \cdot 0.011 = 0.08.$$

Тоді, достовірність цього коефіцієнта регресії визначається наступним чином:

$$t_R = \frac{6.4}{0.08} = \underline{\underline{80.0}}, \quad \nu = 1618 - 2 = 1616,$$

$t_{st}$={2.0−2.6−3.3}

Теоретично можлива максимальна похибка при прогнозі генерального параметра і довірчі межі:

$$\Delta = t_m = 2{,}0 * 0{,}08 = 0{,}16.$$
$$R_{y/x} = = +6{,}4 \pm 0{,}16 = \{6{,}24 - 6{,}56\}.$$

Таким чином, очікуємо, що при збільшенні (або зменшенні) обхвату грудей на 1 см вага коней збільшиться (або зменшиться) у середньому на R = +6,4 кг за умови гарантованого мінімуму зміни +6,24 кг та можливому максимумі +6,56 кг, якщо враховувати зміни ознак в обидві сторони від їх середнього значення.

Коефіцієнт прямолінійної регресії показує, наскільки від свого середнього відхиляється друга ознака, якщо перша ознака від свого середнього відхилилася на одиницю виміру. Це можна виразити такою формулою:

$$(X_2-\mu_2)=R_{2/1}\,(X_1-\mu_1)$$

Позначаючи $X_1$ через $x$, $X_2$ через $y$, $R_{1/2}$ через $b$ і зробивши необхідні перетворення цього виразу можна отримати **робочу формулу прямолінійної регресії:**

$$y=a+bx$$

$$\left\{\begin{array}{l} a=\mu_y-b\mu_x \\ b=R_{y/x} \end{array}\right\}.$$

За цією формулою, знаючи значення $x$ (аргумент), можна визначити значення $y$ (функція) без безпосереднього його виміру: потрібно аргумент $x$ помножити на коефіцієнт регресії та до отриманого результату додати (або відібрати) вільний член $a$. Отже, для попереднього прикладу (визначення ваги коней щодо обхвату грудей) рівняння регресії може бути виведено так:

$$a=\mu_y-R_{y/x}*\mu_x=424-(+6{,}4)\,174=-690,$$
$$b=R_{y/x}=+6{,}4,$$
$$y=a+bx=-690+6{,}4x=6{,}4x-690.$$

Тобто, щоб визначити (без зважування живої ваги коня) за цим способом, треба обхват грудей коня помножити на постійний коефіцієнт 6,4 та від отриманого результату відняти постійне число –690.

На основі рівняння прямолінійної регресії можна заздалегідь розрахувати значення функції для кожного значення аргументу. За обхватом грудей можна визначити живу вагу коня. Якщо ці цифри нанести на графік, по осі абсцис якого відкласти через рівні інтервали значення аргументу (обхвату), а по осі ординат − значення функції (ваги), то отримаємо номограму для визначення ваги коней без зважування та без обчислень.

**Помилки елементів рівняння прямолінійної регресії.** У рівнянні простої прямолінійної регресії:

$$y_x=a+bx$$

виникають одразу три **помилки репрезентативності:**

1. Помилка коефіцієнта регресії:

$$s_b=\frac{\sigma_y}{\sigma_x}\cdot\sqrt{\frac{1-r^2}{n-2}}=\frac{\sigma_y}{\sigma_x}\cdot s_r$$

2. Помилка рівняння регресії, або помилка середньої величини функції для кожного значення аргументу:

$$m_{y_x} = \sigma_y \cdot \sqrt{\frac{1-r^2}{n-2}}$$

За даними прикладу маємо:

$$s_{y_x} = 56.8 \cdot 0.011 = 0.62.$$

Отже, максимальна похибка у визначенні рівня точок лінії регресії при першому порозі ймовірності безпомилкових прогнозів ($\beta_1$ = 0,95, $t_1$=2,0) дорівнюватиме:

$$\Delta = t^* s = 2^* 0{,}62 = \pm 1{,}24 \text{ кг}.$$

**3.** Помилка індивідуальних визначень функції:

$$s_y = \sigma_y \sqrt{1-r^2}$$

Згідно з прикладом:

$$s_y = 56.8\sqrt{1-0.89^2} = 26.2.$$

Отже, індивідуальна похибка у визначенні ваги коней за обхватом грудей за знайденою формулою регресії, і, приймаючи перший поріг ймовірності безпомилкових прогнозів ($\beta_1$ = 0,95, $t_1$=2,0), у крайніх випадках не перевищуватиме:

Δ = 2 * 26 = ± 52 кг.

## *Лекція 8. Дискримінантний, кластерний і факторний аналізи*

Виникнення та розвиток класифікаційних типів аналізу тісно пов'язане із дослідженнями психології часів Зігмунда Фрейда. Тривалий час факторний аналіз сприймався виключно як математична модель психологічної теорії інтелекту, а лише починаючи з 50-х років 20 століття одночасно з розробкою математичного обґрунтування факторного аналізу, цей метод став загальнонауковим. На сьогодні, факторний аналіз є важливою складовою будь-якої статистичної комп'ютерної програми і входить до основного інструментарію наук, що мають справу з багатопараметричним описом об'єктів, включаючи соціологію, економіку, біологію, медицину та інші.

Основна ідея факторного аналізу була сформульована ще Френсісом Гальтоном, основоположником вимірювань індивідуальних відмінностей. Вона зводилася до того що, якщо кілька ознак, виміряних у групі індивідів, змінюються узгоджено, то можна припустити існування однієї загальної причини цієї спільної мінливості – **чинника прихованої (латентної), безпосередньо не доступної виміру змінної**. Отже, **факторний аналіз** – це аналіз впливу окремих факторів на кінцевий результат. Під час аналізу звужують коло чинників, щоб зрозуміти, які саме фактори найбільше впливають на результативний показник та оцінити ступінь цього впливу.

Далі у 1901 році Карл Пірсон висуває ідею **«методу головних осей»**, а Чарльз Спірмен, доводячи свою однофакторну концепцію інтелекту, розробляє математичний апарат з метою оцінки власне цього чинника з безлічі вимірів можливостей. У своїй роботі, опублікованій в 1904 році, Спірмен показав, що якщо ряд ознак попарно взаємопов'язуються один з одним і це може бути формуванням **системи лінійних рівнянь**, що пов'язують всі ці ознаки; один загальний фактор **«загальної можливості»** та по одному специфічному фактору **«спеціальних можливостей»** для кожної змінної. Натомість, Річард Стоун вперше пропонує багатофакторний аналіз для опису численних виміряних можливостей з меншим числом загальних факторів інтелекту, що є лінійною комбінацією цих вихідних можливостей. З 1950-х років з появою комп'ютерів факторний аналіз починає широко використовуватися в психології під час розробки тестів, обґрунтування структурних теорій інтелекту та особистості. При цьому, дослідник починає з безлічі виміряних емпіричних показників, які за допомогою факторного аналізу групуються за факторами (тобто властивостями, що вивчаються). Фактори отримують інтерпретацію щодо змінних, що входять до їх складу, потім відбираються найвагоміші показники цих факторів, відсіваються малозначущі змінні, обчислюються значення факторів для піддослідних і зіставляються із зовнішніми емпіричними показниками досліджуваних властивостей.

Надалі, з розвитком математичного забезпечення факторного аналізу та накопичення досвіду його використання насамперед у психології, завдання факторного аналізу узагальнюється. Як загальнонауковий метод, факторний

аналіз стає засобом для заміни набору корелюючих вимірів істотно меншим числом нових змінних (чинників). При цьому **основними вимогами** є:

| |
|---|
| мінімальна втрата інформації, що міститься у вихідних даних |
| можливість уявлення (інтерпретації) факторів через вихідні змінні |

Таким чином, **головною метою факторного аналізу** є зменшення розмірності вихідних даних з метою їхнього економного опису за умови мінімальних втрат вихідної інформації. **Результатом факторного аналізу** є перехід від безлічі вихідних змінних до значно меншої кількості нових змінних – факторів. Чинник при цьому інтерпретується як причина спільної мінливості декількох вихідних змінних. Водночас, якщо виходити з припущення, що **кореляцію** можна пояснити **впливом прихованих причин – чинників**, то основне призначення **факторного аналізу** – **аналіз кореляцій безлічі ознак.**

В основі факторного аналізу лежить побудова факторної моделі, в якій можна відстежити взаємозв'язок різних чинників із результативним показником. Таких моделей існує велика кількість – **загальна модель факторного аналізу, модель головного компонента, модель специфікації, структурного рівня тощо.** Здійснюється факторний аналіз також за допомогою дослідження різних методів і прийомів. Найчастіше це **дослідницький факторний аналіз (EFA), підтверджуючий факторний аналіз (CFA), метод головних компонент, прийом ланцюгових підстановок, відносних різниць** тощо.

**Загальна модель факторного аналізу** – передбачає існування загального і специфічних факторів, які безпосередньо впливають на змінні. *Наприклад,* за цією моделлю виявляють окремий чи загальний фактор, що впливає на прибуток/наростання маси, витрати фінансові чи біологічні, обсяги продажів/енергія проростання тощо. Також саме таку модель застосовують для оцінки кредитних ризиків та стратегічного планування.

**Модель аналізу факторних завантажень** – визначає, які фактори найбільше впливають на певний результат або змінний показник. *Наприклад,* дослідник може використовувати цю модель при формуванні робочої гіпотези наростання вегетативної маси рослин, щоб врахувати вплив стресорів на ресурси, природні чинники, зміни витрат на майбутній приріст та інші показники. У економіці така модель найчастіше застосовується у банках, коли фінансові менеджери оцінюють кредитні ризики позичальників, щоб розрахувати, як їхній дохід чи кредитна історія впливають на ймовірність неплатежів.

**Модель головних компонент (аналіз основних компонентів, PCA)** – метод, який використовують для спрощення складних даних та виявлення найвагоміших факторів, що впливають на результат. *Наприклад,* директор підприємства може застосовувати PCA для виявлення чинників, які впливають на зміну прибутку компанії. Для цього він має зібрати дані щодо прибутку з фінансових звітів та врахувати можливі впливові фактори: витрати, обсяги продажів, податкові навантаження, процентні ставки та інші показники, а потім застосувати PCA до цих даних, щоб визначити, які фактори найсильніше впливають на зміну прибутку.

Існують також два методи факторного аналізу – **дослідницький та підтверджуючий.** Ці два різних підходи використовують для відокремлення та розуміння прихованих факторів, які можуть впливати на зміни даних. **Дослідницький факторний аналіз** (Exploratory Factor Analysis, EFA) — використовується для виявлення неявних факторів, які можуть бути причинами змін у даних, що визначає структуру факторів у наявних даних. **Підтверджуючий факторний аналіз** (Confirmatory Factor Analysis, CFA) — використовують для підтвердження або перевірки факторів, які було розроблено за допомогою EFA. Ці методи застосовують на різних етапах аналізу даних. Дослідницький аналіз необхідний для побудови структури даних і початкової моделі, у той час, як підтверджуючий дозволяє перевірити цю модель, наскільки вона відповідає даним.

Також існує метод факторного аналізу – **прийом ланцюгових підстановок або метод послідовного вилучення факторів.** Його використовують для розрахунку ступеня впливу окремих факторів на кількісний показник. Суть прийому в тому, що кожен звітний показник послідовно замінюють на базисний, а всі інші показники розглядаються як незмінні. Така підстановка дозволяє виявити рівень впливу фактору на сукупний кількісний показник, а кількість змін залежить від кількості факторів, які на нього впливають. Але у цього метода є недолік – залежно від обраного порядку заміни факторів, результати факторної підстановки можуть мати різні значення. Тому, при цьому аналізі, висуваючи на перший план значущість впливу факторів, важливо не нехтувати точністю даних.

Отже, основний плюс факторного підходу в тому, що він спрощує аналіз даних – зменшує кількість цих даних, звужує коло чинників, зберігаючи при цьому важливу інформацію. До того ж, застосування факторного аналізу надає можливість: виявляти приховані взаємозв'язки між змінними, що дозволяє краще розуміти динаміку змін даних; створювати моделі та прогнозувати ризики; виявляти вузькі місця в роботі підприємства; впливати на кількісні результати.

Універсальність застосування та простота розрахунків зумовлює часте застосування факторного аналізу у фінансах. Однак він також має свої недоліки:

- **вибір факторів може бути суб’єктивним або неправильним** – кількість факторів надто велика або надто мала, через що аналіз

буде неточним. До того ж неправильний вибір може призвести до переоцінки чи недооцінки важливих факторів;

- **лінійність моделі** – більшість моделей факторного аналізу передбачають лінійні зв'язки між факторами та змінними, але у реальних ситуаціях взаємозв'язки можуть бути складнішими;
- **залежність від даних** – результати факторного аналізу сильно залежать від якості та обсягу вхідних даних, тому наявність помилок, пропусків чи інших проблем у даних може призвести до неточних результатів.
- Інтерпретація факторів може бути складною, особливо коли фактори є абстрактними або неочевидними з точки зору практики.

Тож **факторний аналіз** – це універсальний інструмент, який можна успішно застосовувати для розв’язання різних завдань: як для розуміння складних зв'язків кількісних даних для підвищення продуктивності, так і для операційного та стратегічного управління діяльністю.

**Дискримінантний аналіз** – різновид багатовимірного аналізу, призначений для вирішення задач розпізнавання образів. Використовується для прийняття рішення про те, які змінні розділяють (тобто «дискримінують») певні масиви даних (так звані «групи»). Це альтернатива множинного регресійного аналізу для випадків, коли залежна змінна є некількісною (номінативною) змінною. Водночас, дискримінантний аналіз вирішує, по суті, ті ж завдання, що й **множинний регресійний аналіз (МРА):** передбачення значень «залежної» змінної, у даному випадку – категорій номінативної ознаки; визначення того, які «незалежні» змінні найкраще підходять для такого передбачення. Структури вихідних даних для дискримінантного та множинного регресійного аналізу практично ідентичні:

| **№** | $\mathbf{X_1}$ | $\mathbf{X_2}$ | **. . .** | $\mathbf{X_P}$ | **Y** |
|---|---|---|---|---|---|
| **1** | $x_{11}$ | $x_{12}$ | . . . | $x_{1P}$ | $Y_1$ |
| **2** | $x_{21}$ | $x_{22}$ | . . . | $x_{2P}$ | $Y_2$ |
| **. . .** | . . . | . . . | . . . | . . . | . . . |
| ***N*** | $X_{N1}$ | $X_{N2}$ | . . . | $X_{NP}$ | $Y_N$ |

Рядки цієї таблиці відповідають об'єктам (випробуваним), а стовпці – змінним. Змінні *xiXp* представлені у кількісній шкалі, а відмінність вихідних даних для дискримінантного та множинного регресійного методів полягає лише в тому, що є «залежною» змінною Y: для МРА вона є кількісною, а для дискримінантного аналізу – номінативною (класифікуючою) змінною.

У той самий час дискримінантний аналіз можна визначити як метод класифікації, оскільки «залежна» змінна – номінативна, тобто вона класифікує піддослідних на групи, відповідно до різних її градацій. У цьому сенсі вихідними даними для дискримінантного аналізу є група об'єктів (випробувань), розділена на класи G так, що кожен об'єкт віднесений до **одного і тільки одного класу** (градації номінативної змінної). Припускається при цьому, що деякі об'єкти, які не віднесені до якогось із цих класів є «невідомими». Для кожного з об'єктів є дані по Р-кількісним ознаками, однаковими для цих об'єктів. Ці кількісні ознаки називаються **дискримінантними змінними.** Завданнями дискримінаційного аналізу є: визначення вирішальних правил, що дозволяють за значеннями дискримінантних змінних віднести кожен об'єкт (у тому числі і «невідомий») до одного з відомих класів; а також визначення «ваги» кожної дискримінантної змінної для поділу об'єктів на класи. Тобто, дискримінантний аналіз **дозволяє вирішити дві групи проблем**:

**1. Інтерпретувати різницю між класами** – відповісти на питання наскільки добре можна відрізнити один клас від іншого, використовуючи даний набір змінних; які з цих змінних найістотніші для розрізнення класів. Але подібне завдання вирішує і дисперсійний аналіз.

**2. Класифікувати об'єкти**, або віднести кожен об'єкт до одного з класів, виходячи лише з значень дискримінантних змінних. Завдання класифікації пов'язане з отриманням за даними про «відомі» об'єкти дискримінантних функцій «розв'язуючих правил», що дозволяють за значеннями дискримінантних змінних віднести з певною ймовірністю кожен об'єкт до одного з класів.

У розв'язанні задачі класифікації дискримінантний аналіз не є замінним іншими методами. Часто дискримінантний аналіз називають ще «класифікацією з навчанням» чи «розпізнавання образів». У першому випадку припускають, що у такому випадку дослідник «вчиться» класифікувати «невідомі» об'єкти за дискримінантними змінними, використовуючи дані про «відомі» об'єкти. У другому випадку під «образом» об'єкта мається на увазі сукупність відповідних йому значень дискримінантних змінних. І дискримінантний аналіз дозволяє розпізнати образ «нового» об'єкта шляхом віднесення його до відомого класу об'єктів.

**Основні методи дискримінантного аналізу:**

- Лінійний дискримінант Фішера
- Канонічний чи лінійний дискримінантний аналіз (Linear Discriminant Analysis, LDA)
- Логістична регресія.

**Лінійний дискримінантний аналіз** (Linear Discriminant Analysis, LDA), нормальний дискримінантний аналіз (Normal Discriminant Analysis, NDA) або аналіз дискримінантних функцій (Discriminant Function Analysis) є узагальненням **лінійного дискримінанта Фішера** – методу, що використовується для розпізнавання образів та машинному навчанні для пошуку лінійної комбінації ознак, яка описує або поділяє два і більше класів

чи подій. Комбінація, що вийшла, може бути використана як лінійний класифікатор, або, частіше, для зниження розмірності перед класифікацією.

Не дивлячись на певні риси схожості, дисперсійний аналіз використовує якісні незалежні змінні та безперервну залежну змінну, в той час як дискримінантний аналіз має безперервні незалежні змінні та якісну залежну змінну (тобто мітку класу). LDA тісно пов'язаний також з **методом головних компонент** (МГК, Principal Component Analysis, PCA) і факторним аналізом тим, що вони шукають лінійні комбінації змінних, які краще пояснюють дані. LDA намагається моделювати різницю між класами даних, а PCA, з іншого боку, не бере до уваги будь-яку різницю в класах, тоді як факторний аналіз будує комбінації ознак, спираючись швидше на відмінності, а не на подібність.

Дискримінантний аналіз відрізняється також від факторного аналізу тим, що не є незалежною технікою – для його роботи має бути визначена різниця між незалежними змінними та залежними змінними (останні також називаються критеріальними змінними). Також LDA працює, коли виміри, зроблені на незалежних змінних кожного спостереження, є безперервними величинами, а якщо спостереження це якісні незалежні змінні, допустимою еквівалентною технікою є **дискримінантний аналіз відповідностей.**

Дискримінантний аналіз працює шляхом створення однієї чи більше лінійної комбінацій предикторов, отримуючи нову приховану змінну кожної функції. Ці функції називаються **дискримінантними функціями**. Число можливих функцій дорівнює або Ng-1, де Ng = кількості груп, або p (числу предикторов), залежно від того, яке з чисел менше. Перша створена функція максимізує різницю між групами цієї функції. Друга функція максимізує різницю за цією функцією, але не повинна корелювати з попередньою функцією. Процес продовжується створенням послідовності функцій із вимогою, щоб нова функція не корелювала з усіма попередніми.

Якщо дана група j з безліччю Rj вибіркового простору, є дискримінантне правило, таке, що якщо $x \in Rj$, то $x \in j$. Дискримінантний аналіз тоді знаходить «допустимі» області множин Rj для мінімізації помилки класифікації, тому призводить до високого відсотку класифікації.

Кожна функція супроводжується дискримінантною оцінкою визначення, наскільки оптимально вона передбачає належність групі.

- **Коефіцієнти структурної кореляції:** Кореляція між кожним предиктором та дискримінантною оцінкою для кожної функції. Це повна кореляція.
- **Нормовані коефіцієнти:** Вклад кожного предиктора в кожну функцію, тому це є приватною кореляцією. Показує відносну важливість кожного предиктора як внесок у приналежність до групи для кожної функції.
- **Функції від центроїдів групи:** Середні дискримінантні оцінки кожної змінної кожної функції. Чим далі один від одного знаходяться середні, тим меншою буде помилка при класифікації.

Водночас, потрібно враховувати правила дискримінанту:

- **Метод максимальної правдоподібності:** Призначає x групі, що максимізує (групову) густину популяції
- **Правило дискримінанта Байєса:** Призначає x групі, що максимізує πifi(x), де πi представляє апріорну ймовірність класифікації та fi (x) представляє густину популяції.
- **Правило лінійного дискримінанта Фішера:** Максимізує відношення між $SS_{між}$ і $SS_{всередині}$, і знаходить лінійну комбінацію предикторов для передбачення групи.

**Канонічний дискримінантний аналіз** (Canonical discriminant analysis, CDA) знаходить осі (k − 1 канонічних координат, де k – число класів), які найкраще поділяють категорії. Ці лінійні функції не корелюють і визначають, в результаті, оптимальне k − 1 мірний простір через n-мірну хмару даних, які краще розділяють k груп.

**Терміни лінійний дискримінант Фішера та LDA** часто використовують як рівнозначні, наприклад, нормальний розподіл класів або однаковість коваріації класів. Припустимо, що два класи спостережень мають середні μ→0, μ→1 та коваріацій Σ0, Σ1. Тоді лінійна комбінація ознак w→x→ матиме середні w→μ→i та дисперсії w→TΣiw→для i=0,1. Фішер визначав поділ між цими двома розподілами як відношення дисперсії між класами та дисперсії всередині класів:

$$S = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{(\vec{w}\cdot\vec{\mu}_1 - \vec{w}\cdot\vec{\mu}_0)^2}{\vec{w}^T\Sigma_1\vec{w} + \vec{w}^T\Sigma_0\vec{w}} = \frac{(\vec{w}\cdot(\vec{\mu}_1 - \vec{\mu}_0))^2}{\vec{w}^T(\Sigma_0 + \Sigma_1)\vec{w}}$$

Ця дія є, у певному сенсі, мірою відношення сигнал/шум для розмітки класу. Можна показати, що максимальний поділ буде, коли:

$$\vec{w} \propto (\Sigma_0 + \Sigma_1)^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$$

Якщо припущення LDA виконуються, вищенаведена рівність еквівалентна **LDA**.

Варто відмітити, що вектор w→ є нормаллю дискримінантної гіперплощини. Як приклад, у двовимірній задачі пряма, що найкраще розділяє дві групи, є перпендикуляром до w→. У загальному випадку точки даних, які поділяють, проектуються на w→ Потім вибирається граничне значення, яке найкраще розділяє дані, виходячи з одномірного розподілу. Немає загального правила для вибору порога. Однак, якщо проекції точок з обох класів виявляють приблизно той самий розподіл, хорошим вибором буде гіперплощина між проекціями двох середніх, w→μ→0 та w→μ→1. У цьому випадку параметр c у пороговій умові w→x→>c може бути знайдено:

$$c = \vec{w}\cdot\frac{1}{2}(\vec{\mu}_0 + \vec{\mu}_1) = \frac{1}{2}\vec{\mu}_1^T\Sigma_1^{-1}\vec{\mu}_1 - \frac{1}{2}\vec{\mu}_0^T\Sigma_0^{-1}\vec{\mu}_0.$$

**Логістична регресія чи логіт-модель (logit model)** – статистична модель, яка використовується для прогнозування ймовірності виникнення деякої події шляхом її порівняння з логістичною кривою. Ця регресія видає дані у вигляді ймовірності бінарного події (1 чи 0) (рис. 8.1).

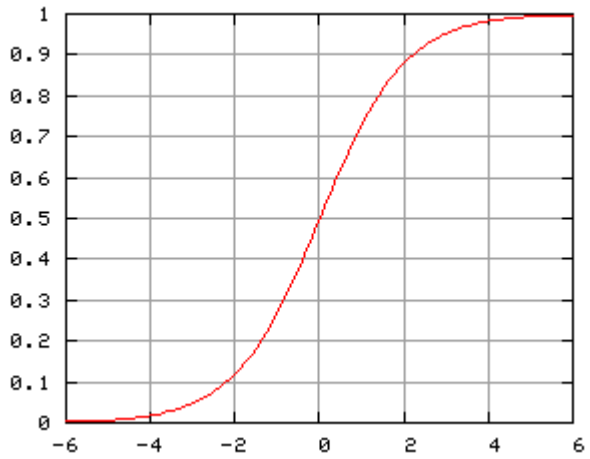



**Рис. 8.1** Приклад логістичної фукції

Логістична регресія застосовується для прогнозування ймовірності виникнення деякої події за значенням множини ознак. Для цього вводиться так звана залежна змінна y, що приймає лише одне з двох значень – як правило, це числа 0 (подія не відбулася) і 1 (подія відбулася), і безліч незалежних змінних (також названих ознаками, предикторами або регресорами) – речових x1,x2….xn, на основі значень яких потрібно обчислити ймовірність прийняття того чи іншого значення залежної змінної. Як і у випадку лінійної регресії, для простоти запису вводиться фіктивна ознака x0=1. Ця модель часто застосовується для вирішення завдань класифікації, об’єкт x можна віднести до класу y=1, якщо передбачена моделлю можливість $P\{y=1|x\}>0,5$, і до класу y=0 в іншому випадку. Отримані при цьому правила класифікації є **лінійними класифікаторами**. Тобто, *наприклад*, дослідник у галузі освіти може досліджувати, які змінні відносять випускника середньої школи до однієї з трьох категорій: вступник до коледжу, вступник у професійну школу, що відмовляється від подальшої освіти. Або, медик може реєструвати різні змінні, які стосуються стану хворого, щоб з'ясувати, які змінні краще показують, що пацієнт, ймовірно, одужав повністю, частково чи зовсім не одужав.

**Кластерний аналіз** (Data clustering) – аналіз, задача якого розбиття заданої вибірки об'єктів (ситуацій) на підмножини, які називаються кластерами, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Завдання кластеризації належить до статистичної обробки, а також до широкого класу завдань некерованого навчання.

**Кластерний аналіз** – це не якийсь один алгоритм, а загальна задача, для розв'язання якої використовуються різні підходи. Зокрема, алгоритми побудови кластерів можуть суттєво відрізнятись у розумінні того, що

відносити в один кластер і як їх ефективно шукати. Серед популярних концепцій кластерів є групи з елементами, які утворюються ґрунтуючись на відстані між ними, на щільності ділянок у просторі даних, інтервалах або на конкретних статистичних розподілах. Тому кластеризація може бути сформульована як задача **багатокритеріальної оптимізації**. Відповідний алгоритм кластеризації та вибору параметрів (включаючи такі параметри, як функція відстані, порогове значення щільності або кількість очікуваних кластерів) залежать від конкретного набору даних та мети використання результатів. Кластерний аналіз як такий є не автоматизованим завданням, а ітераційним процесом виявлення знань або інтерактивної багатокритеріальної оптимізації, який містить спроби та невдачі. Часто доводиться змінювати процес опрацювання даних та параметри моделі поки не буде отримано з результат з заданими властивостями.

Кластерний аналіз вирішує завдання побудови класифікації, тобто **поділу вихідної множини об'єктів на групи (класи, кластери).** При цьому передбачається, що у дослідника немає вихідних припущень ні про склад класів, ні про їхню відмінність один від одного. Приступаючи до кластерного аналізу, дослідник має у своєму розпорядженні лише інформацію про характеристики (ознаки) для об'єктів, що дозволяє судити про подібність (відмінність) об'єктів, або лише даними про їх попарну подібність (відмінність). У літературі найчастіше зустрічаються синоніми кластерного аналізу: **автоматична класифікація, таксономічний аналіз, аналіз образів (без навчання).** Незважаючи на те, що кластерний аналіз відомий відносно давно (вперше викладено Тріоном у 1939 році), поширення ця група методів набула істотно пізніше, ніж інші багатовимірні методи. Лише після публікації книги «Початки чисельної таксономії» біологами Р. Сокел і П. Сніт в 1963 починають з'являтися перші дослідження з використанням цього методу. Проте, досі науці відомі лише поодинокі випадки вдалого застосування кластерного аналізу, попри його виняткову простоту. Викликає здивування наполегливість, з якою у психології використовують для вирішення простого завдання класифікації (об'єктів, ознак) такий складний метод, як факторний аналіз. Разом з тим, кластерний аналіз не лише набагато простіше і наочніше вирішує це завдання, а й має безперечну перевагу: результат його застосування не пов'язаний із втратою навіть частини вихідної інформації про відмінності об'єктів чи кореляцію ознак.

Кластерний аналіз походить з антропології, де він був започаткований Драйвером і Крьобером у 1932 році. У психологію він був введений Зубіним у 1938 році і Робертом Тріоном у 1939 р. Став відомий завдяки використанню Кеттелем для класифікації теорії ознак в психології особистості, починаючи з 1943 року.

Варіанти кластерного аналізу – це безліч простих обчислювальних процедур, що використовуються для класифікації об'єктів. **Класифікація об'єктів** – це групування їх в класи так, щоб об'єкти в кожному класі були схожими один на одного більше, ніж на об'єкти з інших класів. Точніше визначення кластерного аналізу – це процедура впорядкування об'єктів у

порівняно однорідні класи на основі попарного порівняння цих об'єктів за попередньо визначеними та виміряними критеріями.

Існує безліч варіантів кластерного аналізу, але найширше використовуються методи, об'єднані загальною **назвою ієрархічний кластерний аналіз** (Hierarchical Cluster Analysis) – це комбінаторна процедура, що має простий та наочний результат. Можна вказати ряд завдань, при вирішенні яких кластерний аналіз є ефективнішим, ніж інші багатовимірні методи:

- розбиття сукупності випробуваних на групи за виміряними ознаками з метою подальшої перевірки причин міжгрупових відмінностей за зовнішніми критеріями, *наприклад*, перевірка гіпотез про те, чи проявляються типологічні відмінності між випробуваними за виміряними ознаками;
- застосування кластерного аналізу як значно простішого та наочного аналога факторного аналізу, коли ставиться лише завдання угруповання ознак на основі їх кореляції;
- класифікація об'єктів на основі безпосередніх оцінок відмінностей між ними (наприклад, дослідження соціальної структури колективу за даними соціометрії – за виявленими міжособистісними перевагами).

Незважаючи на відмінність цілей проведення кластерного аналізу, можна виділити загальну його послідовність його щодо самостійних кроків, які відіграють істотну роль прикладному дослідженні:

- Відбір вибірки для кластеризації.
- Визначення множини характеристик, по яких будуть оцінюватися об'єкти у вибірці.
- Обчислення значень тієї чи іншої міри схожості між об'єктами.
- Застосування одного з методів кластерного аналізу для створення груп схожих об'єктів.
- Перевірка достовірності результатів кластеризації.

Якщо кластерному аналізу передує факторний аналіз, то вибірка не потребує коректування – викладені вимоги виконуються автоматично самою процедурою факторного моделювання. В іншому випадку вибірку потрібно коректувати.

**Визначення безлічі змінних**, якими відрізнятимуться об'єкти кластеризації – **для випробуваних** – це набір виміряних ознак, для об'єктів, що оцінюються, – суб'єкти оцінки, для ознак – випробувані. Якщо як вихідні дані передбачається використовувати результати попарного порівняння об'єктів, необхідно чітко визначити критерії цього порівняння випробуваними (експертами).

**Визначення міри різниці** між об'єктами кластеризації, **це перша проблема,** яка є специфічною для методів аналізу відмінностей: багатовимірного шкалювання та кластерного аналізу.

**Вибір та застосування методу класифікації** для створення груп подібних об'єктів. **Це друга та центральна проблема** кластерного аналізу, вагомість якої пов'язана з тим, що різні методи кластеризації породжують різні угруповання для тих самих даних. Хоча аналіз і полягає у виявленні структури, в процесі кластеризації структура привноситься в дані, і ця привнесена структура може не збігатися з реальною.

Важливим етапом є перевірка достовірності розбиття на класи – це останній етап який завжди необхідний, наприклад, для виявлення соціальної структури групи. Проте слід пам'ятати, що кластерний аналіз завжди розіб'є сукупність об'єктів на класи, незалежно від того, чи існують вони насправді. Тому марно доводити суттєвість розбиття на класи, наприклад, на підставі достовірності різниці між класами за ознаками, включеними до аналізу. Зазвичай перевіряють стійкість угруповання на повторній ідентичній вибірці об'єктів. Значимість розбиття перевіряють за зовнішніми критеріями – ознаками, які не увійшли до аналізу.

Оскільки поняття «кластеру» не може бути точно визначено, то це є однією з причин чому існує так багато різних методів кластеризації. Але є і спільна риса – це об'єднання схожих об'єктів у групи. Однак, різні дослідники використовують різні моделі кластерів і для кожної з цих моделей можуть бути застосовані різні алгоритми. Поняття кластера, які отримуються у різних алгоритмах, різняться властивостями. Розуміння цих «кластерних моделей» є ключовим для розуміння відмінностей між різними алгоритмами. Типовими кластерними моделями є:

- **Моделі зв'язності.** *Наприклад,* ієрархічна кластеризація або таксономія будуються на основі відстані між вузлами.
- **Центроїдні моделі.** *Наприклад,* метод K-середніх (K-means) представляє кожен кластер єдиним усередненим вектором.
- **Статистичні моделі.** Кластери будуються ґрунтуючись на статистичних розподілах. Таких як багатовимірний нормальний розподіл з допомогою EM-алгоритму.
- **Моделі засновані на щільності.** *Наприклад*, в DBSCAN і в OPTICS кластери визначаються як зв'язані області відповідної щільності у просторі даних.
- **Групові моделі.** Деякі алгоритми не забезпечують вдосконалену модель для своїх результатів, а просто описують групування об'єктів.
- **Графові моделі.** Поняття кліки (така підмножина вершин, в якій кожна пара вершин з'єднана ребром) у графі слугує прототипом кластеру. Пом'якшення вимоги до повної зв'язності (тобто, частина ребер може бути відсутня) призводить до поняття відомого як квазі-кліка. Вони будуються алгоритмом HCS.
- **Нейронні моделі.** Найвідомішою моделлю нейронної мережі з некерованим навчанням є нейронна мережа Кохонена. Ці моделі, як правило, можна охарактеризувати як схожі на одну або подібні

якійсь з наведених вище моделей, включаючи моделі у підпросторах, коли нейронні мережі реалізують метод головних компонент або аналіз незалежних компонент.

**«Кластеризацією»** зазвичай вважають такий набір кластерів, які містять усі об'єкти набору даних. Додатково, можна розглянути відношення між кластерами. *Наприклад,* ієрархію вкладеності кластерів один у одного. Грубо можна виділити такі типи кластеризації:

1. Жорстка кластеризація. Кожен об'єкт або належить кластеру або ні.
2. М'яка кластеризація (також нечітка кластеризація). Кожен об'єкт належить кожному кластеру до певної міри. Наприклад, це ймовірність належності кластеру.

Серед них виділяють декілька внутрішніх:

- Жорстке розбиття на кластери. Кожен об'єкт належить рівно одному кластеру.
- Жорстке розбиття на кластери з викидами. Об'єкт може не належати жодному кластеру і розглядається як викид.
- Кластери з перетином. Об'єкт може належати більш ніж одному кластеру.
- Ієрархічна кластеризація. Якщо об'єкт належить нащадку, то він також належить і предку.
- Підпросторова кластеризація. Хоч кластери і можуть перетинатись, проте в межах визначеного підпростору кластери не перетинаються. Для прикладу дивись SUBCLU.

Вхідними даними кластерного аналізу є **набір об'єктів.** Залежно від способу представлення цих об'єктів розрізняють такі типи вхідних даних:

- Вектор характеристик. Кожен об'єкт описується набором своїх характеристик; ці характеристики можуть бути числовими або нечисловими.
- Матриця відстаней. Кожен об'єкт описується відстанями до всіх інших об'єктів вибірки.

Кластерний аналіз висуває наступні вимоги до даних:

- Об'єкти не повинні корелювати між собою.
- Об'єкти мають бути безрозмірними.
- Розподіл об'єктів має бути близьким до нормального.
- Об'єкти повинні відповідати вимозі стійкості, під якою розуміється відсутність впливу на їх значення випадкових чинників.
- Вибірка повинна бути однорідна.

Рішення задачі кластеризації принципове неоднозначне, і цьому є декілька причин:

1. Не існує однозначно якнайкращого критерію якості кластеризації. Відомий цілий ряд евристичних критеріїв, а також ряд алгоритмів, що не мають чітко вираженого критерію, але

здійснюють достатньо розумну кластеризацію «по побудові». Всі вони можуть давати різні результати.

2. Число кластерів, як правило, невідоме заздалегідь і встановлюється відповідно до деякого суб'єктивного критерію.
3. Результат кластеризації істотно залежить від метрики, вибір якої, як правило, також суб'єктивний і визначається експертом.

**Результатом кластеризації** є групи об'єктів, об'єднані за певною характеристикою чи характеристиками. Однак ці результати можуть бути інтерпретовані по-різному. Зокрема, при аналізі результатів соціологічних досліджень рекомендується здійснювати аналіз ієрархічними методами, наприклад методом Уорда, при якому всередині кластерів оптимізується мінімальна дисперсія і в результаті створюються кластери приблизно рівних розмірів. Як міра відмінності між кластерами використовується квадратична евклідова відстань, що сприяє збільшенню контрастності кластерів.

Тепер виникає питання стійкості знайденого кластерного рішення. По суті, перевірка стійкості кластеризації зводиться до перевірки її достовірності. Тут існує емпіричне правило – стійка типологія зберігається при зміні методів кластеризації. Результати ієрархічного кластерного аналізу можна перевіряти ітеративним кластерним аналізом методом k-середніх. Якщо при порівнянні групи збігаються більше, ніж на 70 % (понад 2/3 збігів), то кластерне рішення приймається за основу.

Перевірити адекватність рішення, не вдаючись до допомоги інших видів аналізу, не можна. Принаймні, теоретично ця проблема не вирішена. Деякі додаткові методи перевірки стійкості відкидаються з певних причин:

- Кофенетична кореляція – не рекомендується і обмежена у використанні.
- Тести значущості (дисперсійний аналіз) – завжди дають значущий результат.
- Метод повторних випадкових вибірок – не доводить правильність рішення.
- Тести значущості для зовнішніх ознак – придатні тільки для повторних вимірювань.
- Методи Монте-Карло – дуже складні і доступні тільки досвідченим математикам.

***Використана література***

*Базова*

1. Wayne W. Daniel, Chad L. Cross Biostatistics: A Foundation for Analysis in the Health Sciences, 11th Edition. – Wiley, 2018. – 720 p.
2. Harvey Motulsky Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking, 3rd edition. – Oxford University Press. – 2018. – 576 p.
3. Горошко М.П. Біометрія / М.П. Горошко, С.І. Миклуш, П.Г. Хомюк. – Львів, Камула, 2004. – 285 с.

*Допоміжна*

1. Triola Mark, Jason Roy Biostatistics for the Biological and Health Sciences, 2nd edition. – Pearson Education, 2018. – 420 p.
2. Атраментова Л. О. Біометрія : підруч. для студ. вищ. навч. закладів / Л. О. Атраментова, О. М. Утєвська. – Харків : Ранок, 2007. – 176 с.
3. B. Burt Gerstman Basic Biostatistics: Statistics for Public Health Practice, 2nd edition. – Jones & Bartlett Learning, 2014. – 648 p.
4. Горкавий В. К. Статистика : підручник / В. К. Горкавий. – К. : Аграрна освіта, 2009. – 511 с.
5. Jan Leps Biostatistics with R, 1st edition. – Cambridge University Press, 2020. – 384 p.
6. Калінін М. І. Біометрія [Електронний ресурс]: підручник для студ. вузів біол. і еколог. напрямів / М. І. Калінін, В. В. Єлісеев. – Режим доступу http://lib.chdu.edu.ua/index.php?m=1&b=3
7. Max Kuhn, Kjell Johnson Applied Predictive Modeling [eBook]: Book for Mathematics and Statistics // Springer New York, NY.
https://doi.org/10.1007/978-1-4614-6849-3
https://link.springer.com/search?facet-content-type=%22Book%22&package=11649&facet-start-year=2013&facet-end-year=2013