

ЛЕКЦІЯ 6

ТЕМА: Основи математичного статистичного аналізу результатів досліджень.

1. Основні поняття математичної статистики.
2. Основні завдання математичної статистики.
3. *Аналіз варіаційних рядів кількісної та якісної мінливості.*

Самостійна робота: Облік врожаю.

1. Основні поняття математичної статистики

Результати досліджень аналізують методами математичної статистики.

Математична статистика — це розділ математики, який розглядає методи систематизації, обробки та використання статистичних даних для формування кінцевих висновків і пропозицій. В математичній статистиці використовують відповідні поняття, терміни та символи. Об'єктам агрономічних досліджень — рослинам, їх групам та навколишньому середовищу — властиве таке явище, як мінливість. Мінливість властивостей рослин та їх середовища називають *варіюванням*. Ступінь варіювання, виражений математично, називається *варіацією*.

Великі вибірки (кількість об'єктів більше 30). Якщо 1000 насінин однієї культури чинавіль одного сорту, що являє собою генеральну сукупність, висіяти в однакових умовах, середовища, рослини будуть розрізнятися ростом, зовнішнім виглядом, масою, врожаєм тощо. Для визначення середньої висоти рослин чи середньої кількості зерен на великій площі треба було б терміново, протягом кількох годин вимірювати тисячі рослин і рахувати десятки тисяч зерен, що практично неможливо. Якщо подібні обліки треба зробити по кількох сортах, у різних варіантах, у кількох повторностях, то дослідження неможливо виконати у стислі строки (користуються теорією ймовірностей).

Теорія ймовірностей — це наука про закономірності масових випадкових явищ. Згідно з цією теорією замість суцільного обліку великої генеральної сукупності для вивчення відбирають певну частину, за якою і оцінюють всю генеральну сукупність. Отже, за ймовірностями одних випадкових подій визначають ймовірність інших подій, пов'язаних з ними.

Для аналізу результатів польових дослідів користуються рівнем надійної імовірності 0,95, який записують символом $P_{0,95}$, а для більш точних аналізів — рівнем 0,99, котрий записують символом $P_{0,99}$. Рівень надійної імовірності $P_{0,95}$ означає, що на ньому рівні дослідник, стверджуючи чи заперечуючи в досліді певне явище чи положення, ризикує помилитись 5 разів на 100 випадків. На рівні $P_{0,99}$ ризик помилитися становить 1 раз на 100.

Емпіричні розподіли, які мають у дослідях, не завжди такі симетричні, як теоретичні. У них може бути скривленість (**скошеність**) вправо або вліво, тобто асиметричність. Іноді у скривлених розподілів може бути не одна, а кілька вершин. Це свідчить про неоднорідність вибірки. Наприклад, у вибірку потрапили рослини іншого сорту або варіанта, який істотно відрізняється від першого.

Малі вибірки. Поряд з великими вибірками ($n > 30$) в агрономічних дослідженнях часто користуються вибірками, коли n дорівнює 4—8 або 10—12 варіантам у досліді. При цьому число членів ряду становить менше 30 і їх називають малими вибірками, на які не можна переносити закони великих чисел. Для малих вибірок застосовують розподіл ймовірностей Стюдента, який називають *законом малих вибірок*. Для них запропонований критерій Стюдента, що позначається символом t .

Критерій достовірності (істотності). *Критерій* — це показник, який дає змогу робити висновок про надійність висновків відносно статистичної гіпотези. Найчастіше користуються нульовою гіпотезою (H_0). Нульова гіпотеза — припущення про відсутність реальної різниці між фактичними спостереженнями (x_f) і тими, що передбачались теоретично (x_T). Для перевірки статистичних гіпотез користуються критеріями достовірності. Критерії достовірності бувають параметричні та непараметричні.

Параметричні критерії достовірності застосовують лише для нормального розподілу, найчастіше це критерії Стюдента (t) та Фішера (F). Критерій достовірності Стюдента розраховують за формулою

$$t = d / S_d$$

Де d — різниця середніх арифметичних ($x_1 - x_2$), а S_d — похибка різниці.

Розрахункове (фактичне) значення критерію t порівнюють з теоретичними значеннями на певних рівнях імовірності.

Критерій достовірності Фішера (F) прямо пропорційний дисперсії варіантів і обернено пропорційний дисперсії залишку похибки. Його фактичне значення порівнюють з теоретичним, яке знаходять у дод. 3.

2. Основні завдання математичної статистики

Одним з основних завдань наукових досліджень, де застосовують математичну статистику, є *планування дослідів*. Суть його така: запланований дослід повинен мати достатню кількість варіантів і повторностей, всі варіанти на початку дослідів мають перебувати в однакових умовах

навколишнього середовища, треба правильно підібрати метод статистичної обробки результатів дослід. Саме ці питання і вирішують математичним плануванням.

Завданням математичної статистики є також і те, яку *запланованому і закладеному досліді відбирати об'єкти для досліджень*, щоб вони об'єктивно характеризували вплив факторів, які вивчаються. Тут ідеться про використання методу рандомізації при відбиранні зразків ґрунту чи рослин у досліді.

Не менш важливим є вирішення питання про те, *скільки взяти цих зразків для досліджень*, тобто як оптимізувати об'єм вибірки.

У процесі попередньої обробки даних вирішують також *завдання з відновлення дат*, які випадають з дослід, і для цього користуються певними формулами.

Іноді одержані у досліді дані того самого варіанта дуже розрізняються і постає питання про *перевірку їх належності до певного варіаційного ряду*, а якщо вони сумнівні, то вибракувати. З цією метою користуються або критерієм тау (т).

У кожному проведеному досліді *визначають достовірність різниць між середніми арифметичними досліджуваних вибірок із застосуванням критеріїв достовірності t, F, а також найменшої істотної різниці (НІР)*.

У багатьох дослідженнях, де вивчають залежності між показниками середовища і рослиною або між різними показниками самої рослини (наприклад, урожай і якість), *обчислюють коефіцієнти кореляції та кореляційні відношення*.

Питання прогнозування, або знаходження невідомих показників за допомогою інших відомих, вирішують у регресійних аналізах після складання рівнянь регресії для лінійних чи криволінійних залежностей.

Майже у всіх дослідженнях виникає *питання про точність досліджень*. Для цього спочатку обчислюють значення відносної похибки (5 %), а за її значенням роблять висновки про точність проведених досліджень.

Головною особливістю математичної статистики є те, що її можна застосовувати лише у методично правильно спланованих і проведених дослідях. Якщо методики досліджень недодержували, ніяка винахідливість математика не зможе виправити похибок дослід. Такі досліді треба бракувати.

3. Аналіз варіаційних рядів кількісної та якісної мінливості

Розрізняють мінливість кількісну і якісну. До кількісної мінливості належать об'єкти, які мають масу, розмір, об'єм або їх можна лічити поштучно. Розрізняють непереривну та переривну кількісну мінливість. До *непереривної мінливості* належать ті об'єкти, які виражають в основному дробовими числами,— це маса, розмір і об'єм досліджуваних об'єктів: маса

врожаю або плоду, довжина і висота рослин, площа листя, об'єм коріння чи бульби. До *переривної мінливості* належать об'єкти, які обліковують поштучно: кількість колосів, листків, коренеплодів, рослин тощо.

Конкретний хід аналізу варіаційних рядів кількісної мінливості залежить від об'єму вибірки. Для малих і великих вибірок обчислюють такі основні статистичні характеристики: середню арифметичну($\bar{x}$); дисперсію; стандартне відхилення; похибку середньої арифметичної; коефіцієнт варіації— V ; відносну похибку. Кожній середній арифметичній дають інтервальнооцінку.

Малі вибірки. Прикладом малих вибірок може бути кількість повторностей, яка найчастіше становить 3—6.

Для малих вибірок обчислюють такі статистичні характеристики: середні арифметичні, дисперсії, стандартні відхилення, коефіцієнти варіації, похибки вибірових середніх, відносні похибки та ін.

Середня арифметична проста — $\bar{x}$. Для обчислення статистичних характеристик використаємо, наприклад, дані обліку кількості листя кукурудзи на полі з рівномірним та дуже нерівномірним розсіюванням добрив.

Варіюючі ознаки (кількість листя) позначають знаком X , а кількість повторень — n . Середню арифметичну просту обчислюють за формулою

$$\bar{X} = \Sigma X / n$$

де $\bar{X}$ — середня арифметична для варіанта; ΣX — сума варіюючих ознак варіанта по повторностях, у даному випадку — кількість листя на одній рослині кукурудзи; n — кількість повторень.

Для поля з нерівномірним внесенням добрива середня арифметична

$$\bar{X}_2 = \Sigma X/n = (16+7+20+17)/4=15$$

Для варіанта з рівномірним внесенням добрив середня арифметична

$$\bar{X}_1 = \Sigma X/n = (14+18+13+15)/4=15$$

Отже, середні арифметичні однакові, але розмах варіації різний. При рівномірному внесенні добрив = 5, а при дуже нерівномірному =13, тобто кількість листя сильніше варіює при нерівномірному внесенні добрив, коли створюється строкатість у родючості ґрунту.

Середня арифметична є основною статистичною характеристикою кожного варіаційного ряду (всі інші характеристики лише пояснюють головну).

Дисперсія (S²). При повторних обліках на одних і тих самих полях, тобто на одних генеральних сукупностях, розмах варіювання може бути неоднаковим. Отже, він є менш характерним показником варіації. Повніше характеризує варіаційні ряди дисперсія (S²) — середній квадрат відхилень кожного члена варіаційного ряду ($X_1, X_2, \dots, X_n$) від їх середньої арифметичної. Дисперсію обчислюють за формулою

$$S^2 = \Sigma (X - \bar{X})^2 / n - 1$$

X – варіюючи ознака; $\bar{X}$ – середня арифметична;

$n-1$ – число всіх вимірювань без одиниці, або v
можна використовувати метод статистичної обробки для дослідів з неповним числом дат.

Вибір методу статистичної обробки даних. Якщо дані не викликають сумніву, обчислюють середні арифметичні для кожного варіанта, вибирають метод статистичної обробки і виконують відповідний аналіз, наприклад дисперсійний. Вибір дисперсійного аналізу залежить від методу розміщення варіантів у польових дослідів.

Для дослідів, розміщених методом рандомізованих повторень, використовують дисперсійний аналіз рандомізованих повторень. Досліди, розміщені методом латинського квадрата, обробляють дисперсійним аналізом латинського квадрата, а для дослідів, розміщених методом латинського прямокутника, застосовують дисперсійний аналіз латинського прямокутника.

Багатофакторні досліди, які розміщують методом рандомізованих повторень або розщеплених ділянок, обробляють відповідними методами дисперсійного аналізу — рандомізованих повторень чи розщеплених ділянок.

У дослідів, розміщених методом повної рандомізації, а також у вегетаційних лабораторних та дослідів, в яких вивчають зберігання картоплі, буряків, тощо, повторень не організовують і дані обробляють дисперсійним аналізом повної рандомізації.

Недисперсійні методи статистичної обробки застосовують для дослідів, у яких не використовують випадковий метод розміщення варіантів, наприклад стандартний та систематичний методи. Досліди, розміщені стандартним методом, обробляють різницеvim методом, а розміщені систематично — дробовим.

Показники якісної мінливості обробляють визначенням достовірності різниць між частками наявності ознак за допомогою критерію Стюдента (t).

ЛЕКЦІЯ 7

ТЕМА: Статистичні методи перевірки гіпотез.

1. Дисперсійний аналіз.
2. Не дисперсійні методи статистичної обробки дослідів.
3. Кореляційний та регресійний аналізи.
4. Коваріаційний аналіз.

Самостійна робота: Використання комп'ютерної техніки для планування досліджень, проведення обліків і спостережень, створення баз даних, їх аналізу та інтерпретації.

1. Дисперсійний аналіз.

Дисперсійний аналіз є найдосконалішим методом статистичної обробки даних. Проте, його застосовують лише для дослідів, які розміщені методом рендомізації (випадково).

У польовому досліді, розміщеному методом рендомізованих повторень, урожай змінюється залежно від:

- варіантів;
- повторостей;
- випадкових причин (умови навколишнього середовища, індивідуальні мінливості саміх рослин).

Р.Фішер виразив ці зміни сумами квадратів таких розсіювань:

- варіантів - C_v
- повторень - C_p
- похибки - C_z

їх сума і є сумою квадратів загального розсіювання (C_y).

Тоді $C_y = C_v + C_p + C_z$

Для кожного розсіювання обчислюють число ступенів свободи (v):

- варіантів - $v_v = l-1$;
- повторень - $v_p = n-1$;
- похибки - $v_z = (l-1) \cdot (n-1)$;

загального розсіювання - $N-1$, де $N = l \cdot n$, (l -кількість варіантів, n -кількість повторень)

Дисперсія — це розсіювання даних дослідів і розчленування загального

варіювання вро-жаю чи інших показників на складові частини. Звідси і назва методу — дисперсійний аналіз. Найбільш застосовувані дисперсія варіантів та дисперсія помилки, яку ще називають дисперсією залишку. Співвідношення цих двох дисперсій є тим основним критерієм, який дає змогу дати загальну оцінку достовірності різниць між середніми арифметичними або загальну оцінку достовірності дослідів. Цей критерій позначають першою літерою прізвища автора дисперсійного аналізу Фішера

Якщо критерій Фішера фактичний (обчислений) дорівнює критерію теоретичному

$$F=(S^2v)/(S^2z)$$

Найбільш застосовувані дисперсія варіантів (S^2v) та дисперсія помилки (S^2z).

Крім того визначення критерію Фішера викликає питання про визначення НІР.

Дисперсійний аналіз є найдосконалішим методом статистичної обробки даних. Його переваги полягають у виділенні із загального варіювання його компонентів, розрахуванні узагальненої похибки всього дослідів (Е) на основі більшої кількості спостережень, ніж для індивідуальних похибок окремих пар варіантів у недисперсійних методах.

Дисперсійний аналіз досить ефективний для багатофакторних дослідів, оскільки дає змогу визначити достовірність не лише дії факторів окремо, а і їх взаємодії.

Висновок про точність усього дослідів роблять наприкінці дисперсійного аналізу на основі числового значення відносної похибки $S_x\%$, яку визначають за формулою:

$$S_x\% = (E-100)/X_N$$

Е-узагальнена похибка дослідів; X_N - середня арифметична всього дослідів.

Без обчислення похибки дослідів дисперсійний аналіз вважається незакінченим, а висновки неповними.

2. Недисперсійні методи статистичної обробки дослідів.

Застосовуються для статистичної обробки результатів, одержаних в дослідів з нерендомізованим розміщенням варіантів. До них відносять *дробовий та різницевий методи*.

Дробовий метод статистичної обробки результатів дослідів.

Метод запозичений із біометрії, де мають справу з великою кількістю повторних спостережень. Таким умовам часто відповідають результати польових і лабораторних дослідів більш-менш однорідної сукупності, варіювання якої обумовлено випадковими факторами.

Різницевий метод статистичної обробки результатів дослідів.

Застосовують для дослідів, розміщених стандартними методами (ямб, дактиль) методами. Найчастіше їх використовують у сортовивченні, а також в інших дослідів за умов сильного варіювання родючості ґрунту.

3. Кореляційний та регресійний аналізи.

Методи статистичного аналізу, які розглядалися вище, дають можливість вивчати окремі ознаки або властивості незалежно, наче ізольовано від інших. В багатьох агрономічних дослідженнях дуже важливо вияснити залежність між двома або декількома ознаками, встановити їх взаємний зв'язок. Але в таких дослідженнях рідко мають справу з точними і визначеними функціональними зв'язками, коли кожному значенню однієї величини відповідає строго визначене значення іншої величини. Частіше зустрічаються такі співвідношення між змінними, коли кожному значенню ознаки X відповідає не одна, а безліч можливих значень ознаки Y. Такі зв'язки pojawiaються лише при масовому визначенні ознак і на відміну від функціональних називаються *схоластичними (вірогідними) або кореляційними*. Класифікація кореляцій (взаємне співвідношення, залежність показників, явищ тощо) поділяються за:

- 1) напрямом;
- 2) формою;
- 3) силою;
- 4) кількістю зв'язків.

1) За напрямом вони бувають прямі та зворотні.

Пряма кореляція-коли із збільшенням однієї ознаки (X) інша ознака (Y) також збільшується.

Приклад: із збільшенням довжини колосу кількість зерен в ньому також збільшується; збільшення довжини листя призводить до збільшення їх площі.

Зворотна кореляція - із збільшенням однієї ознаки (X) інша ознака (Y) зменшується.

Приклад: при більшій забур'яненості посівів зменшується врожайність польових культур; збільшення доз застосовуваного інсектициду зменшує кількість шкідників на полі; надмірне розростання коренеплодів цукрових буряків призводить до зниження їх цукристості.

При вивченні кореляційних зв'язків виникає два основних питання - про силу зв'язків та їх форму. Для вимірів сили і форми зв'язків використовують спеціальні статистичні методи, які називаються кореляцією і регресією. Кореляція та регресія - спеціальні статистичні методи для визначення сили і форми зв'язків.

1) За формою кореляції поділяють:

-прямолінійні- із збільшенням одних ознак (X) відповідно збільшується інші ознаки (Y).

Приклад: при збільшенні маси бульб картоплі чи коренеплодів буряків збільшуються їх розміри; при збільшенні кількості зерен у колосі збільшується його довжина.

Ступінь кореляційних взаємозв'язків виражається числом, яке називається *коефіцієнтом кореляції* (r).

- *криволінійні* - має місце, коли значення X та Y змінюються спочатку в одному напрямі, а потім у протилежному.

Приклад: при постійному зростанні градацій певного фактору X (азотні добрива, вологість ґрунту) врожай Y спочатку зростає, потім стає стабільним, а після подальшого збільшення ознаки X ознака Y починає зменшуватись. Такі зв'язки виражають кореляційним відношенням, яке позначають r (грецьке «ета»).

2) За кількістю зв'язків кореляція буває:

- *простою* (досліджується зв'язок між двома ознаками);
- *множинною* (досліджується зв'язок між трьома і більшою кількістю ознак). Приклад множинної кореляції: залежність урожайності одночасно від доз добрив, норм зрошення, норм висіву, глибини загортання насіння.

3) За силою зв'язків кореляція може бути:

- повною;
- сильною;
- середньою.
- слабою;
- її не може бути зовсім.

Значення коефіцієнту кореляції відображає силу зв'язку. Показник змінюється від -1 до +1. При значенні $r \pm 1$ кореляція перетворюється на функціональну. Як що $r=0$ між X та Y немає зв'язку, але криволінійна залежність може існувати.

При: - $r < 0,3$ зв'язок слабкий;

- $r = 0,3 \dots 0,7$ зв'язок середній;
- $r > 0,7$ зв'язок сильний.

Додатне значення (+) r вказує на пряму позитивну, а від'ємне (-) на обернену (негативну) кореляцію. Низький коефіцієнт кореляції не завжди визначає відсутність зв'язку, тоді може бути криволінійна залежність.

Коефіцієнт кореляції указує на напрям і ступінь взаємозв'язку, але не дає можливості зробити висновок про те, як кількісно змінюється функціональна ознака (Y) при зміні факторіальної ознаки (X) на одиницю виміру. Тобто в таких випадках на допомогу дослідників приходить *регресійний аналіз*.

Регресійний аналіз - дає можливість передбачити значення функціональної ознаки, за заданим значенням факторіальної, тому що регресія вказує на ступінь зміни ознаки Y при зміні на одиницю ознаки X.

Наприклад: із зміною довжини колоса (X) на 1 см кількість зерен у ньому (Y) збільшується на 7 штук.

Після кореляційних та регресійних аналізів складають рівняння регресії, які використовують для обчислення невідомого показника за відомим.

4. Коваріаційний аналіз.

У дослідах з багаторічними рослинами іноді на дослідних ділянках дерева значно розрізняються за силою росту або врожайністю на початку досліду. Такі рослини, як правило, ще більше будуть різнитися між собою і наприкінці досліду. Тому оцінка ефективності варіантів без поправок на

початковий стан рослин не буде об'єктивною, У таких випадках необхідно знайти співвідношення між варіюванням початкового стану показника, наприклад урожаю X_i кінцевого V . Коваріаційний аналіз використовується також у тих випадках, якщо в процесі дослідження випадають деякі рослини, зіпсовані шкідниками, морозами, уражені хворобами. Проте якщо в сортовивченні сильне ураження хворобами або морозами є особливістю сорту або в агротехнічних дослідженнях випадання рослин сталося під впливом високих доз добрив або гербіцидів, коваріаційний аналіз не застосовується.

Коваріаційний аналіз — сукупність методів математичної статистики, що відносяться до аналізу моделей залежності середнього значення деякої випадкової величини Y одночасно від набору (основних) якісних факторів F_i (супутніх) кількісних факторів X . Фактори задають поєднання умов, при яких були отримані спостереження X, Y, X, Y , і описуються за допомогою індикаторних змінних, причому серед супутніх і індикаторних змінних можуть бути як випадкові, так і не випадкові (контрольовані в експерименті).

Комплекс дисперсійного, кореляційного, та регресійного аналізів з метою зведення фактичних середніх ряду V до повної вирівняності умов дослідження за рядом X називається коваріаційним аналізом. У математичній статистиці коваріація (cov) — це середній добуток відхилень двох змінних X та Y від їх середніх.

