ЛЕКЦІЯ 23-24
МЕТОДИ АНАЛІЗУ ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ. ПОБУДОВА МАТЕМАТИЧНИХ МОДЕЛЕЙ

Аналіз експериментальних даних

В дослідженнях для обробки експериментальних даних найбільш широко застосовуються такі методи математичної статистики, як дисперсійний, кореляційний і регресійний аналіз.

Дисперсійний аналіз

Дисперсійний аналіз – основна задача – визначення впливу різних факторів на мінливість ознаки, яка вивчається. Наприклад урожай в польових умовах, успішність студентів. Загальне варіювання (мінливість) - S_y можна розчленувати на три основні частини:

варіювання варіантів - S_v;

варіювання повторів - S_p;

випадкові варіювання - S_t

$$S_y = S_v + S_p + S_t \quad (1)$$

Особливостями дисперсійного аналізу є такі положення:

1. Замість середніх для окремих варіантів досліду обчислюється одна загальна середня арифметична для всього досліду в цілому.

2. Замість індивідуальних помилок середніх кожного варіанта досліду обчислюють одну усереднину похибку загальної середньої, яку використовують для оцінки розрізнювання варіантів.

3. Середню похибку досліду знаходять шляхом розкладання загальної дисперсії всіх даних досліду на складові частини, які характеризують варіювання, яке пов’язане з факторами, які вивчаються в досліді, і варіювання випадкове, яке обумовлене різноманітним випадковим впливом зовнішніх умов на мінливість при знаків і властивостей, які вивчаються.

Визначення випадкового варіювання часто є основною задачею дисперсійного аналізу. Воно дає можливість визначити помилку досліду і найменшу суттєву різницю (Н С Р), тобто ту мінімальну різницю між середніми, яка в даному експерименті є суттєвою

$$HCP = t \cdot S_d$$

де t – критерій Стьюдента для прийнятого рівня значущості і числа ступенів волі залишкової дисперсії (береться з таблиці).

S_d – похибка різниці обчислюється за формулою

$$S_d = \sqrt{\frac{2S_z^2}{n}} = 1.41 S_{\bar{x}} \quad (2)$$

де n – число, що повторюється в порівняльних варіантах;

S_z^2 - залишковий середній квадрат (дисперсія помилок);

$S_{\bar{x}}$ - узагальнена помилка середньої

Вибираємо 5% рівень значущості, що означає, що похибка може повторитися 5 раз із 100.

$$HCP_{05} = t_{05} \cdot S_d \quad (3)$$

Кореляційний і регресійний аналіз

Якщо необхідно визначити залежність між двома або декількома признаками і встановити їх взаємний зв’язок використовують кореляції і регресії. Теорія кореляції вивчає взаємозв’язок між величинами, які досліджуються. Діалектичний підхід до вивчення природи і суспільства вимагає розглядати явища у взємозв’язку і в неперервному змінюванні. Теорія кореляції дозволяє виразити ці взаємозв’ки у кількісній формі.

Найбільш простим видом зв’язку між величинами є функціональна залежність, коли кожному значенню однієї величини відповідає одне конкретно визначене значення другої величини.

До функціональних зв’зків відноситься наприклад, залежність між об’ємом води W, часом t і використанням Q:

$$Q = \frac{W}{t} \quad (4)$$

Якщо змінна величина у змінюється в залежності від іншої змінної х, але на зміну у впливає багато інших факторів, врахувати які інколи не в змозі, то тоді кожному значенню х відповідає декілька значень у. Такі зв’зки називаються кореляційними, або зв’язок між змінними величинами х і у називається кореляційним, якщо різним значенням однієї із них (х) відповідають групові середні другої (у) або навпаки. В таких випадках одна величина розглядається як незалежна змінна і називається аргументом (х), а друга – залежна змінна і називається функцією (у). Загальний вигляд рівняння кореляційного зв’язку у=f(х), де х - аргумент, у – функція.

При графічному зображенні статистичного звя’зку часто точки розміщують так, що можна провести ряд ліній різноманітного типу.

Після встановлення форми зв’язку і її типу визначають її тісноту. В якості числового показника зв’язку простої лінійної кореляції використовують коефіцієнт кореляції

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \quad (5)$$

де $(x - \bar{x})$ і $(y - \bar{y})$ - відхилення значень х і у від своїх середніх $\bar{x}$ і $\bar{y}$ в п порівнювальних парах.

Стандартну похибку коефіцієнта кореляції визначають з рівняння

$$S_r = \sqrt{\frac{1-r^2}{n-2}} \quad (6)$$

r – коефіцієнт кореляції; п – число пар значень, за якими обчислений коефіцієнт кореляції. Значення коефіцієнта кореляції записується разом з його похибкою у вигляді $r = \pm S_r$. Критерій суттєвого коефіцієнта кореляції t обчислюють з рівняння

$$t_r = \frac{r}{S_r} \text{ або } t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (7)$$

Зіставлення фактичного і теоретичного (табличного) значень t при числі ступеню волі п-2 дає можливість оцінити суттєвість r при тому чи іншому рівню значущості.

Якщо $t_{r,факт} \geq t_{теор}$, то кореляційний зв'язок існує, а якщо $t_{r,факт} \leq t_{теор}$ - не існує.

Поряд з коефіцієнтом кореляції для характеристики зв'язку між двома ознаками використовують коефіцієнт детермінації d_{yx}, який чисельно рівний квадрату коефіцієнта кореляції:

$$d_{yx} = r^2 \quad (8)$$

Коефіцієнт детермінації показує частину тих змін, які у залежності, яку вивчають обумовлені факторіальними ознаками і дають більш чітке уявлення про ступінь спряження ознак. Наприклад, якщо коефіцієнт кореляції рівний 0,20 – 0,30, то коефіцієнт детермінації $d_{yx} = 0.04 - 0.09$ тобто тільки 4-9% всіх вимірів однієї ознаки пов'язані із змінами другої. При $r = 0.5 - 0.6$ число зв'язків збільшується до 25-30% і тільки при $r = 0.95$ біля 97% зміна результативної ознаки пов'язано із змінами факторіального.

Кореляційне відношення обчислюється

$$\eta_{yx} = \sqrt{\frac{S_v}{S_y}} \quad (9)$$

де η – кореляційне відношення; S_v – сума квадратів відхилення за варіантами;

S_y – загальна сума квадратів.

Кореляційне відношення використовується для оцінки криволінійної форми зв'зку між ознаками і має додатній знак, змінюється від 0 до 1.

При малому числі спостережень кореляційне відношення обчислюється:

$$\eta_{xy} = \sqrt{\frac{\sum f(\bar{y}_x - y)^2}{\sum f(y - \bar{y})^2}} \quad (10)$$

де $\sum f(y - \bar{y})^2$ - сума квадратів відхилень групових і середніх y_x від загальної середньої y (групове варіювання), яка характеризує ту частину варіювання ознаки y, яка пов'язана з мінливістю ознаки x.

$\sum f(y - \bar{y})^2$ - сума квадратів різниці між кожним значенням і загальною середньою $\bar{y}$, яка характеризує загальне варіювання ознаки y.

Похибка S_η і критерій істотного кореляційного відношення обчислюється за рівнянням:

$$S_\eta = \sqrt{\frac{1-\eta^2}{n-2}}; \quad t_\eta = \frac{\eta}{S_\eta} \quad (11)$$

Фактичне значення t_η порівнюють з теоретичним, який приймається для вибраного рівня значущості при числі ступенів волі $n-2$ з таблиці. Якщо $t_\eta \geq t_{0.5}$, то кореляційне відношення суттєве.

Квадрат кореляційного відношення називають індексом детермінації:

$$d_{yf(x)} = \eta^2_{yz} = \frac{\sum f(\bar{y}_x - \bar{y})^2}{\sum f(y - y)^2} \quad (12)$$

Він показує ту долю варіювання ознаки y, яка обумовлена змінами ознаки x.

Обчисливши коефіцієнт кореляції можна отримати загальну уяву про спряження ознак які вивчаються.

Регресійний аналіз – наукове дослідження закономірностей між явищами, які залежать від багатьох факторів. Мета його – відшукати рівняння лінії, яка найбільш точно виражає залежність однієї ознаки від іншої. За формою регресія може бути прямолінійною і криволінійною, а за характером – простою, коли змінювання вислідної ознаки відбувається під зміною однієї факторіальної ознаки, і множинною, коли зміна обумовлена декількома факторіальними ознаками.

Регресивний аналіз дозволяє передбачити можливість зміни однієї ознаки на основі відомих змін другої шляхом розрахунку емпіричних формул, які показують, що зв'язок між ними існує.

При лінійній регресії залежність між ознаками виражається коефіцієнтом регресії, який показує в якому напрямку і на яку величину змінюється одна ознака при зміні другої на одиницю виміру.

Обчислюється коефіцієнт регресії за рівняннями:

$$b_{yx} = r\frac{S_y}{S_x}; \quad b_{yx} = r\frac{S_x}{S_y}. \quad (13)$$

Де r - коефіцієнт кореляції;

S_x і S_y - середні квадратичні відхилення;

x і y вивчаються у рядах.

Коефіцієнти регресії мають знак коефіцієнта кореляції:

$$b_{yx}b_{xy} = r^2 \quad (14)$$

Ця властивість використовується для перевірки чи правильно обчислений коефіцієнт регресії.

Похибку коефіцієнтів регресії обчислюють за рівнянням:

$$S_{byx} = S_r \frac{S_y}{S_x} \quad \text{і} \quad S^b_{yx} = S_r \frac{S_x}{S_y}. \quad (15)$$

Критерій суттєвості коефіцієнта регресії дорівнює критерію суттєвості коефіцієнта кореляції, тобто:

$$t_b = \frac{b}{S_b} = t \quad (16)$$

Часто залежність між признаками, які вивчаються буває криволінійною, вона може мати різні форми і описується відповідними рівняннями. В цьому випадку, головна задача регресійного аналізу полягає в тому, щоб по характеру розпределення точок на графіку підібрати аналітичну залежність, яка описує закономірність зміни ознак. Після того, Як аналітична залежність підібрана, необхідно математичними перетвореннями привести її до рівняння прямої лінії, тобто перетворити вихідні дані і обчислити значення параметрів, які входять в аналітичну залежність. Приведення криволінійної залежності до рівняння прямої лінії дозволяє використати прийоми регресійного аналізу.

Приклад.

Техніку приведення кореляційного і регресійного аналізу розглянемо на прикладі для невеликого числа спостережень (x) від змінної (y). x - вологість грунту; y - наліплювання грунту.

1. Розрахунки зручно вести складаючи таку таблицю.

Розрахунки допоміжних величин для обчислення кореляції і регресії y по x.

№	Значення ознаки		x^2	y^2	xy
пари	x (%)	y (г/см^2)			

1	19,9	0,0	396,01	0,00	0,00
2	20,9	0,6	436,81	0,36	12,54
3	26,1	1,1	681,21	1,21	28,71
4	29,4	1,2	864,36	1,44	35,28
5	30,5	1,7	930,25	2,89	51,85
6	40,3	1,7	1624,09	2,89	68,51
7	44,8	2,6	2007,04	6,76	116,48
8	47,8	3,4	2284,84	11,56	162,52
9	55,6	4,2	3091,36	17,64	233,52
10	58,3	5,8	3398,89	33,64	338,14
11	64,5	6,3	4160,25	39,69	406,35
12	76,6	7,3	5867,56	53,29	559,18
Сума	$\sum x = 514.7$	$\sum y = 35.9$	$\sum x^2 = 25742.7$	$\sum y^2 = 171.4$	$\sum xy = 2013.1$

Розв’язання:

2. За даними таблиці обчислюємо шість допоміжних величин: $n = 12$;

$\bar{x} = (\sum x) : n = 514.7 : 12 = 42,89\%;$

$\bar{y} = (\sum y) : n = 35,9 : 12 = 2,99 г / см^2$

$\sum (x - \bar{x})^2 = \sum x^2 - (\sum x)^2 : n = 25742.7 - (514.7)^2 : 12 = 3666,3;$

$\sum (y - \bar{y})^2 = \sum y^2 - (\sum y)^2 : n = 171,4 - (35,9)^2 : 12 = 63,97;$

$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - (\sum x \cdot \sum y) : n = 2013,1 - (514,7 \cdot 35,9) : 12 = 473,3.$

3. Обчислюється коефіцієнт кореляції, регресії і рівняння регресії:

коефіцієнт кореляції

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{473.3}{\sqrt{3666.3 \cdot 63.97}} = 0.977;$$

коефіцієнт регресії x і y

$$b_{yx} = \frac{\sum (x-\bar{x})^2 (y-\bar{y})^2}{\sum (x-\bar{x})^2} = \frac{473.3}{3666.3} = 0.13;$$

Рівняння регресії

$$y = \bar{y} - b_{yx}(x-\bar{x}) = 2.99 + 0.13(x-42.89) = 0.13x - 2.58.$$

Таким чином шукана залежність має вигляд: $y = 0.13x - 2.58.$

4. Визначається похибка і критерій значущості для коефіцієнта кореляції:

Похибка коефіцієнта кореляції

$$S_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.977^2}{12-2}} = 0.067;$$

критерій значущості коефіцієнта кореляції

$$t_r = \frac{r}{S_r} = \frac{0.977}{0.067} = 14.59$$

5. Фактичне значення t_r порівнюється з теоретичним $t_{0.5}$, яке приймається рівним: 8-9 ступенів волі (при $n-2$ - це 10-11 пар спостережень) – 2,3; для 10-14 ступенів волі – 2,2; для 15-24 ступенів волі – 2,1; для 25-100 ступенів волі – 2,0. Кореляція і регресія визначається суттєвою, якщо $t_r \geq t_{0.5}$. В нашому прикладі $t_r \geq t_{0.5}$, так як $14.59 \geq 2.2$. Значить між вологістю грунту і її налипання є суттєвий прямий зв'язок.

6. За отриманим рівнянням регресії обчислюють теоретичне значення y для крайніх величин x (19,9 і 76,6, згідно таблиці)

$y_{x=19.9} = 0.13 \cdot 19.9 - 2.58 = 0.00$;

$y_{x=76.6} = 0.13 \cdot 76.6 - 2.58 = 7.37.$

Знайдені точки ($x = 19.9;\ y = 0.00$ і $x = 76.6;\ y = 7.37$) наносяться на графіці, з'єднуючи їх прямою, маємо теоретичну лінію регресії. Вона показує, що збільшення вологості грунту на 1% відповідає збільшенню налипання на 0,13 г/см^2.

3. ПАРНА РЕГРЕСІЯ

Парна залежність може бути апроксимована прямою лінією, параболою, гіперболою, логарифмічною, степеневою або показниковою функцією,поліномом і інше.

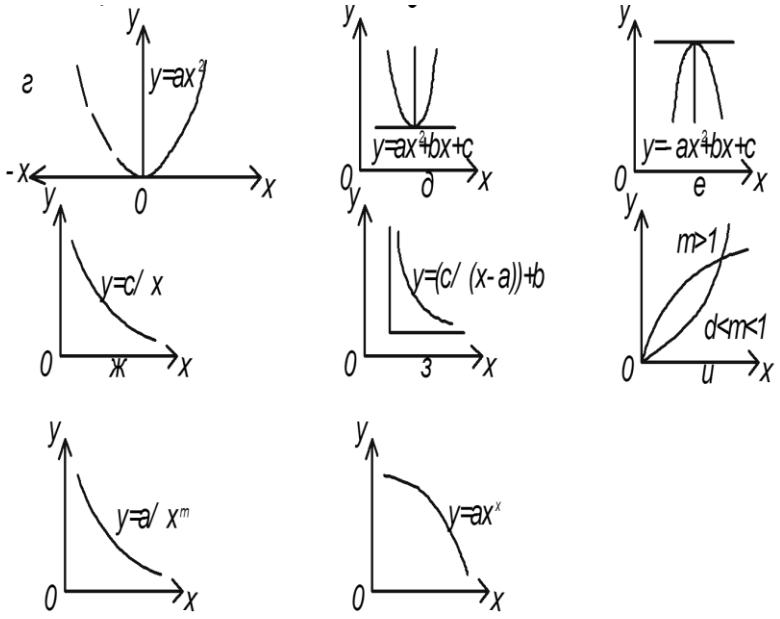

Рис. Вигляди основних ліній різних зв’язків між змінними величинами і їх рівняння.

1. Пряма, яка проходить через початок координат має рівняння $y = ax$ (3,а).

2. Пряма, що не проходить через початок координат має рівняння $y = ax + b$, або $y = -ax + b$. Ці залежності вимагають визначення двох параметрів a і b. (3, б, в).

3. Парабола з вершиною в початку координат і симетрична одній із осей має рівняння $y = ax^2$. Формула один параметр a із зменшенням якого зменшується розхил параболи (рис.3, г).

4. Парабола, симетрична прямій паралельній осі y має рівняння $y = ax^2 + bx + c$. Функція квадратична. У формулі необхідно визначити три параметра: a, b і c (рис.3, д, є).

5. Гіпербола, асимптотично наближається до осей координат, рівняння має вигляд $y = \frac{c}{x}$, необхідно визначити параметр c (рис.3, ж).

6. Гіпербола асимптотично наближається до прямих, паралельних до осей координат, рівняння має вигляд $y = \frac{c}{x-a} + b$. Параметри a і b є координатами точки m. Знак параметра c залежить від розміщення гіперболи по відношенню до асимптот (рис.3, з).

7. Степеневі криві (рис.3, и, к), рівняння $y = ax^m$, де m може бути додатнім, цілим або дробовим.

8. Показникові крива, коли із зростанням однієї величини (x) спостерігається підсилене зростання (y). Рівняння $y = a^x$ (рис.8.3, л).

Двох факторне поле можна апроксимувати, площиною, параболоїдом другого порядку, гіперболоїдом. Для n - змінних фактів зв'язок можна встановити за допомогою n - мірного простору рівняннями другого порядку

$$y = b_0 + \sum_1^n b_i x_i + \sum_1^n b_{ij} x_i x_j + \sum_1^n b_{ij} x_i^2 \quad (17)$$

де y - функція мети багатофакторних змінних;

x_i - незалежні фактори;

b_i - коефіцієнт регресії, що характеризують вплив фактора x_i на функцію мети;

b_{ij} - коефіцієнти, які характеризують подвійний вплив факторів x_i і x_j на функцію мети.

При побудові теоретичної регресійної залежності, оптимальною буде така функція, в якій виконуються умови найменших квадратів $\sum (y_i - \bar{y})^2 = \min$, де y_i - фактичні координати поля; $\bar{y}$ - середнє значення ординати з абсцисою x, обчисленою з рівняння. Після кореляції апроксимують рівнянням прямої. Лінію регресії розраховують з умови найменших квадратів:

$$y = a + bx \quad (18)$$

При цьому крива АВ найкращим чином вирівнює значення постійних коефіцієнтів a і b, тобто коефіцієнтів рівняння регресії. Їх обчислюють за формулами:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \quad (19)$$

$$a = y - bx = \frac{\sum y}{n} - b\frac{\sum x}{n} \quad (20)$$

Критерієм близькості кореляційної залежності між x і y до лінійної функціональної залежності є коефіцієнт парної або просто коефіцієнт кореляції r. Він просто показує ступінь лінійності зв'язку x і y.

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}} \quad (21)$$

де n - число вимірів.

Задовільна тіснота зв'язку при $r \geq 0.5$, добра при $r = 0.8 - 0.85$. Для визначення проценту мінливості шуканої функції y відносно її середнього значення, який визначається мінливістю фактора x, обчислюють коефіцієнт детермінації

$$K_Д = r^2 \quad (22)$$

Рівняння регресії прямої записати таким виразом:

$$y = \overline{y} + r\frac{\delta_y}{\delta_x}(x - \overline{x}) \quad (23)$$