

Практична робота №6

Тема: Кореляційний та регресійний аналізи

6.1 Поняття про кореляцію.

6.2 Кореляційний та регресійний аналіз прямолінійної залежності.

6.3 Аналіз криволінійної залежності.

6.4 Складання рівнянь регресії для криволінійної залежності.

Мета: Навчитись проводити кореляційних та регресійний аналізи прямолінійної та криволінійної залежностей.

6.1 Поняття про кореляцію.

Методи статистичного аналізу, які розглядалися вище, дають можливість вивчати окремі ознаки або властивості незалежно, наче ізольовано від інших.

В багатьох агрономічних дослідженнях дуже важливо виявити 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 появляються лише при масовому вивченні ознак і на відміну від функціональних називаються схоластичними (вірогідними) або кореляційними.

В класифікаційному плані кореляції (взаємне співвідношення, залежність показників, явищ тощо) поділяються за напрямом, формою, силою та кількістю зв'язків.

За напрямом вони бувають прямі та зворотні. Пряма кореляція спостерігається тоді, коли із збільшенням однієї ознаки (X) інша ознака (Y) також збільшується. Приклади прямої кореляції: із збільшенням довжини колоса кількість зерен у ньому також збільшується; збільшення довжини листя призводить до збільшення їх площі; інтенсивніше освітлення рослин посилює фотосинтез тощо. Зворотна кореляція спостерігається, коли із збільшенням однієї ознаки (X) інша ознака (Y) зменшується. Приклад зворотної кореляції: 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 тощо.

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

За формою кореляції поділяють на прямолінійні та криволінійні. При прямолінійній кореляції із збільшенням одних ознак (X) відповідно збільшуються інші ознаки (Y). Наприклад, при збільшенні маси бульб картоплі чи коренеплодів буряків збільшуються їх розміри, при збільшенні кількості зерен у колосі збільшується його довжина та ін.

Криволінійна кореляція має місце, коли значення X та Y змінюються спочатку в одному напрямі, а потім у протилежному. Приклади криволінійної кореляції: при постійному зростанні градацій певного фактору X (азотні або інші добрива, вологість ґрунту тощо) 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 (грецьке «ета»).

За кількістю зв'язків кореляція буває простою, коли досліджується зв'язок між двома ознаками, та множинною, якщо вивчається зв'язок між трьома і більшою кількістю ознак. Приклади простої кореляції наведені вище, а прикладами множинної можуть бути: залежність урожайності одночасно від доз добрив, норм зрошення, норм висіву, глибини загортання насіння тощо; залежність якості врожаю від вище перелічених умов, залежність урожайності від атмосферних опадів, температури повітря, його вологості тощо.

За силою зв'язків кореляція може бути повною, сильною, середньою, слабкою або її може не бути зовсім.

Ступінь кореляційної залежності вимірюється різними показниками. За прямолінійної кореляції ступінь взаємозв'язків виражається числом, яке називається коефіцієнтом кореляції і позначається буквою r , а за криволінійної кореляції - кореляційним відношенням, яке позначається буквою.

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. В таких випадках на допомогу дослідникові приходить регресійний аналіз, за яким можна 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.

6.2 Кореляційний та регресійний аналіз прямолінійної залежності.

Серед переважної більшості криволінійних зв'язків, які зустрічаються в агрономії, деякі з них близькі до прямолінійних і їх зручніше аналізувати з використанням коефіцієнта кореляції.

Кореляційний аналіз. Проведемо аналіз залежності між довжиною 20 окремих листків озимої пшениці та їх площами (табл. 1), визначених на основі індивідуальних вимірів.

Таблиця 1

Обчислення кореляційної залежності між довжиною листя озимої пшениці та його площею

Номери листків (пар)	Довжина листа, см X	Площа листа, см ² Y	X-x	Y-y	(X-x)(Y-y)	(X-x) ²	(Y-y) ²
I	16,1	7,4	-6,4	-7,8	49,92	40,96	60,84
2	17,3	8,7	-5,2	-6,5	33,80	27,04	42,25
3	18,6	10,3	-3,9	-4,9	19,11	15,21	24,01
4	20,0	11,2	-2,5	-4,0	10,00	6,25	16,00
5	21,3	12,9	-1,2	-2,3	2,76	1,44	5,29
6	21,6	13,2	-0,9	-2,0	1,80	0,81	4,00
7	21,8	13,7	-0,1	-1,5	0,15	0,01	2,25
8	22,0	14,1	-0,5	-1,1	0,55	0,25	1,21
9	22,4	14,3	-0,1	-0,9	0,09	0,01	0,81
10	22,8	14,8	0,3	-0,4	-0,12	0,09	0,16
11	23,1	15,2	0,6	0,0	0	0,36	0
12	23,3	16,2	0,8	1,0	0,80	0,64	1
13	23,3	16,7	0,8	1,5	1,20	0,64	2,25
14	23,7	17,0	1,2	1,8	2,16	1,44	3,24
15	24,0	17,4	1,5	2,2	3,30	2,25	4,84
16	24,1	19,2	1,6	3,0	4,80	2,56	9,00
17	25,2	19,3	2,7	4,1	11,07	7,29	16,81
18	26,0	20,3	3,5	5,1	17,85	12,25	26,01
19	26,5	21,4	4,0	6,2	24,80	16,00	38,44
20	26,4	22,3	3,9	7,1	27,69	15,21	50,41
n = 20	x = 22,5	y = 15,2	Σ ≈ 0	Σ ≈ 0	Σ[(X-x)·(Y-y)] = 211,73	Σ(X-x) ² = 150,71	Σ(Y-y) ² = 308,82

Після проведення в таблиці 1 розрахунків далі обчислюємо: коефіцієнт кореляції

$$r = \frac{\sum[(X-x) \cdot (Y-y)]}{\sqrt{\sum(X-x)^2 \cdot \sum(Y-y)^2}} = \frac{211,73}{\sqrt{150,71 \cdot 308,82}} = \frac{211,73}{215,74} = 0,98$$

похибку коефіцієнта кореляції

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0,98^2}{20 - 2}} = \sqrt{\frac{1 - 0,96}{18}} = 0,047$$

критерій достовірності коефіцієнта кореляції

$$t_r = \frac{r}{S_r} = \frac{0,98}{0,047} = 20,85$$

Теоретичне значення критерію Стюдента знаходять за числом ступенів вільності $\nu_r = n - 2 = 20 - 2 = 18$

$$t_{0,95} = 2,1;$$

$$t_{0,99} = 2,88$$

Про силу зв'язку роблять висновок за таким правилом: якщо коефіцієнт кореляції дорівнює одиниці, то зв'язок повний; якщо він становить 0,66-0,99, то зв'язок сильний; якщо він знаходиться в межах 0,33-0,66 - середній; якщо коефіцієнт кореляції менший за 0,33, то зв'язок слабкий.

Оскільки в нашому прикладі $r = 0,98$, то зв'язок між довжиною листа пшениці і його площею сильний.

Про напрям зв'язку висновок роблять залежно від знака при коефіцієнті кореляції: якщо він плюсовий, то кореляція пряма, а якщо мінусовий, то зворотна.

В нашому прикладі кореляція пряма.

Про достовірність зв'язку висновок роблять за правилом: якщо критерій достовірності коефіцієнта кореляції фактичний більший за теоретичні його значення або дорівнює їм, то зв'язок достовірний.

Висновок: оскільки критерій фактичний (t_r) складає 20,85, що значно більше теоретичних значень $t_{0,95}$ (2,1) і $t_{0,99}$ (2,88), то зв'язок між довжиною листа озимої пшениці і його площею достовірний на обох рівнях надійної імовірності.

Регресійний аналіз. Його здійснюють при сильному та достовірному зв'язку і будь-якому напрямі (прямому чи зворотному). Під регресією розуміють зміну результативної ознаки Y (функції) при певній зміні одної або декількох факторіальних (аргументів).

Зв'язок між функцією і аргументом виражають рівнянням регресії, яке має такий вигляд:

$$Y = y + R_{yx}(X - x)$$

Користуючись даними таблиці 42 розрахуємо коефіцієнт регресії R_{yx} , який покаже зміну площі листа озимої пшениці при зміні його довжини на 1 см.

$$R_{yx} = \frac{\sum (X - \bar{x})(Y - \bar{y})}{\sum (X - \bar{x})^2} = \frac{211,73}{150,71} = 1,4 \text{ см}^2.$$

Отже, в нашому прикладі при зміні довжини листка на 1 см його площа

змінюється на $1,4 \text{ см}^2$.

Підставивши значення коефіцієнта регресії у рівняння регресії ми одержимо робоче рівняння, за яким, знаючи довжину листа, можна визначити його площу.

$$Y = y + R_{yx}(X - x) = 15,2 + 1,4 (X - 22,5) = 15,2 + 1,4X - 31,5$$

Таким чином $Y = 1,4X - 16,3$.

Вводячи в це рівняння чисельне значення довжини листа X , яке є середньою величиною від вимірювання 50-100 листків, вираховують площу листа. Її можна визначити для цілої рослини або на певній площі посіву.

Наприклад, після розрахунків довжина листка склала 23,3 см (дванадцята пара у табл. 1), а фактична площа листка при цій довжині становить $16,2 \text{ см}^2$.

Розрахуємо площу листка, взявши його середню довжину 23,3 см, скориставшись виведеним рівнянням регресії

$$Y = 1,4 \cdot 23,3 - 16,3 \approx 16,3 \text{ см}^2$$

Точність прогнозування площі листа озимої пшениці за довжиною її листа розраховують за формулою.

Різниця між розрахунковою площею і фактичною становить $16,3 - 16,2 = 0,1 \text{ см}^2$ або це складає, $(0,1 \cdot 100) : 16,2 = 0,62 \%$ тому точність дослідів дорівнюватиме

$$T \% = 100 - 0,62 = 99,38 \% \approx 99,4 \%$$

Таким чином, точність прогнозування площі листа озимої пшениці за довжиною листка є високою.

6.3 Аналіз криволінійної залежності

Наявність криволінійного зв'язку можна визначити за графіком, коли на ньому одержується вигнута лінія, яка значить, що при зростанні X спочатку Y зростає, а потім зменшується чи навпаки. Це свідчить про криволінійну залежність. Для аналізу цієї залежності користуються не коефіцієнтом кореляції, а кореляційним відношенням η_{xy} або η_{yx} , яке розраховують за формулою:

$$\eta_{yx} = \sqrt{\frac{\sum (X - \bar{x})^2 - \sum (Y - \bar{y})^2}{\sum (Y - \bar{y})^2}},$$

де Y - значення функціональної ознаки; y - середнє значення цієї ознаки.

Обчислимо кореляційне відношення для залежності врожайності зерна озимої пшениці від норм висіву (табл. 2).

Таблиця 2

Урожайність зерна озимої пшениці (Y) залежно від норм висіву (X)

Номери пар (повторність)	Норма висіву, ц/га X	Урожайність, ц/га Y	y_x	$Y - y_x$	$(Y - y_x)^2$	$Y - y$	$(Y - y)^2$
1	3,5	32	33	-1	1	-14	196
2	3,6	34		1	1	-12	144
3	3,7	38	42	-4	16	-8	64
4	3,8	46		4	16	0	0
5	3,9	51	53	-2	4	5	25
6	4,0	55		2	4	9	81
7	4,1	57	55	2	4	11	121
8	4,2	53		-2	4	7	49
9	4,3	50	47	3	9	4	16
10	4,4	44		-3	9	-2	4
		$y = 46$		$\Sigma(Y - y_x) = 0$	$\Sigma(Y - y_x)^2 = 68$	$\Sigma(Y - y) = 0$	$\Sigma(Y - y)^2 = 700$

При десяти варіантах норм висіву їх доцільно розділити на 8 груп, як показано в таблиці, і зробити розрахунки середнього для групи значення Y. Для першої групи $y_x = (32 + 34) : 2 = 33$, а для другої $(38 + 46) : 2 = 42$ і так для інших груп. Значення $Y - y_x$ визначають відніманням від значень Y, які входять до групи, їх середнього:

$$32 - 33 = -1; 34 - 33 = 1; 38 - 42 = -4; 46 - 42 = 4 \text{ і т.д.}$$

Ці відхилення підносять до квадрата, додають і одержують значення $\Sigma(Y - y_x)^2 = 68$. Відхилення обчислюють, віднімаючи від кожного із значень урожайності (Y) середнє значення урожаїв (y): $32 - 46 = -14$; $34 - 46 = -12$ і т.д. Всі одержані числа підносять до квадрата, додають і одержують $\Sigma(Y - y)^2 = 700$. Обидві суми використовують для розрахунку кореляційного відношення η_{yx} :

$$\eta_{yx} = \sqrt{\frac{\Sigma(Y - y)^2 - (Y - y_x)^2}{\Sigma(Y - y)^2}} = \sqrt{\frac{700 - 68}{700}} \approx 0,95$$

Висновок: оскільки кореляційне відношення становить 0,95 і знаходиться в межах 0,66-0,99, то між нормами висіву і врожайністю зерна озимої пшениці зв'язок сильний.

Далі обчислюють похибку кореляційного відношення

$$S_{\eta_{yx}} = \sqrt{\frac{1 - \eta_{yx}^2}{n - 2}} = \sqrt{\frac{1 - 0,095^2}{10 - 2}} = 0,11$$

Для висновків про достовірність зв'язку розраховують критерій достовірності фактичний:

$$t_{\eta} = \frac{\eta_{yx}}{S_{\eta_{yx}}} = \frac{0,95}{0,11} = 8,64.$$

Критерій Стюдента теоретичний знаходять за числом ступенів вільності $v_{\eta} = n - 2 = 10 - 2 = 8$.

Тоді $t_{0,95} = 2,31$, а $t_{0,99} = 3,36$.

Висновок: оскільки критерій Стюдента фактичний $t_{\eta} = 8,64$ більший за $t_{0,95}$ і $t_{0,99}$, то зв'язок достовірний на обох рівнях надійної імовірності.

6.4 Складання рівнянь регресії для криволінійної залежності

В зв'язку з тим, що у наведеному вище прикладі зв'язок сильний і достовірний, використаємо ці дані при складанні рівняння регресії для криволінійної залежності, графічне зображення якої має форму параболи.

Урожайність зерна озимої пшениці на основі норм висіву в нашому прикладі можна розрахувати за формулою

$$Y = y + \frac{\sum(X - x) \cdot Y}{\sum(X - x)^2} \cdot (X - x) + \left[\frac{\sum(X - x)^2 \cdot Y - nCy}{\sum(X - x)^4 - nCy^2} \right] \times [(X - x)^2 - C],$$

де C (коректуючий фактор) $\sum(X - x)^2 : n$.

Щоб за даною формулою зробити розрахунки, складають таблицю 3.

Таблиця 3

Розрахунки вихідних даних для складання квадратичної параболи

Норма висіву, млн./га X	Урожай - ність, ц/га Y	X-x	(X -x) ²	(X - x) ⁴	(X - x)Y	{X -x) ² Y
3,5	32	-0,45	0,2025	0,0410	-14,4	6,480
3,6	34	-0,35	0,1225	0,0150	-11,9	4,165
3,7	38	-0,25	0,0625	0,0039	-9,5	2,375
3,8	46	-0,15	0,0225	0,0005	-6,9	1,035
3,9	51	-0,05	0,0025	0	-2,55	0,128
4,0	55	0,05	0,0025	0	2,75	0,138
4,1	57	0,15	0,0225	0,0005	8,55	1,283

4,2	53	0,25	0,0625	0,0039	13,25	3,313
4,3	50	0,35	0,1225	0,0150	17,50	6,125
4,4	44	0,45	0,2025	0,0410	19,80	8,910
x = 3,95	y = 46,0	$\Sigma(X - x)^2 = 0$	$\Sigma(X - x)^2 = 0,825$	$\Sigma(X - x)^4 = 0,1208$	$\Sigma(X - x)Y = 16,60$	$\Sigma (X-x)^2 Y = 33,95$

Попередньо розраховують коректуючий фактор $C = \frac{\Sigma(X-x)^2}{N} = 0,825 : 10 = 0,0825$

Підставивши його значення як і цифрові значення інших складових вище наведеної формули, одержимо рівняння регресії, за яким можна прогнозувати величину урожайності озимої пшениці (Y) на основі даних про норму висіву насіння (X).

$$Y = 46 + \frac{16,6}{0,825} \cdot (X - 3,95) + \left[\frac{33,95 - 10 \cdot 0,0825 \cdot 46}{0,1208 - 10 \cdot 0,0825^2} \right] \cdot [(X - 3,95)^2 - 0,0825] = 46 + 20,1212X - 79,4787 + \frac{-4}{0,0527} \cdot (X^2 - 7,9X + 15,6025 - 0,0825) = 46 + 20,1212X - 79,4787 - 75,9X^2 + 599,61X - 1184,2298 + 6,2618 = 619,7312X - 75,9X^2 - 1211,4467$$

Щоб перевірити точність прогнозування врожайності зерна озимої пшениці за цим рівнянням регресії, візьмемо норму висіву 4 млн. га (див. шосту стрічку табл. 3) і, підставивши у рівняння, матимемо розрахункову врожайність пшениці даного сорту

$$Y = 619,7312 \cdot 4 - 75,9 \cdot 4^2 - 1211,4467 = 2478,9248 - 1214,4 - 1211,4467 = 53,0781 \approx 53,1 \text{ ц/га}$$

Точність прогнозування урожайності озимої пшениці за нормою висіву розраховують за формулою

$$T \% = \frac{53,1 \cdot 100}{55} = 96,54 \%$$

Отже, точність нашого прогнозування добра, що вказує на можливість практичного використання виведеного рівняння регресії.

Запитання для самоперевірки:

Що таке кореляція і якою вона буває?

Що таке основний показник кореляційного аналізу та його використання?

Для чого проводиться регресійний аналіз?