

Застосування кореляційного та регресійного аналізів у дослідженнях з землеробства

Мета роботи: засвоєння методики кореляційного та регресійного аналізів.

Теорія методу. *Кореляція (англ. correlation) — це статистична залежність між випадковими величинами ознак, що не має чіткого функціонального характеру, вираженого одновекторними змінами цих величин. На відміну від функціональної залежності, за кореляції будь-якій величині однієї ознаки може відповідати декілька величин іншої. Вона відображає закон множини причин і наслідків і є вільною неповною залежністю.*

Метою кореляційного аналізу стає виявлення сили, напрямку та істотності кореляції однієї змінної ознаки від інших.

За формою кореляція може бути лінійною і криволінійною. За кількістю досліджуваних ознак відрізняють її варіанти:

1. парна кореляція, коли досліджується зв'язок між двома ознаками (результативною і факторіальною або двома факторіальними);
2. часткова кореляція визначає залежність між результативною і однією факторіальною ознакою за фіксованої величини інших факторіальних ознак;
3. множинна кореляція, коли на величину однієї результативної ознаки впливають декілька факторіальних.

Математичним критерієм парної лінійної кореляції між двома ознаками x та y слугує **коефіцієнт кореляції** (r), який розраховують за однією із формул:

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

Величина коефіцієнта кореляції набуває модуля в межах від +1 до -1.

Сила кореляційного зв'язку кількісно виражається величиною коефіцієнта кореляції. Чим ближче r до +1 чи -1, тим тісніший прямолінійний кореляційний зв'язок. **Вважають, що коли $r = 0$ – між x і y немає лінійного зв'язку, за $r < 0,3$ кореляційна залежність між ознаками *слабка*, $r = 0,3–0,7$ – *середня* і $r > 0,7$ – *сильна*.**

Під час використання коефіцієнта кореляції слід мати на увазі, що його величина виявляє тільки силу зв'язку. Сам же по собі зв'язок може бути наслідком причинної залежності однієї ознаки від іншої, а може й бути результатом простого випадку. Причинність або випадковість зв'язку, його напрям визначає сам дослідник.

Квадрат коефіцієнта кореляції (r^2) називається коефіцієнтом детермінації і позначається d_{yx} . Він показує частку або відсоток тих вимірювань, які в даному явищі залежать від фактора, що вивчається.

Розвитком кореляційного аналізу є регресійний. Регресія показує кількісну зміну результативної (залежної) ознаки (функції Y) за зміни факторіальної ознаки (аргументу X) на одиницю його виміру. Вона теж є проста, якщо функція досліджується в залежності від одного аргументу ($y = f(x)$), і множинна – від двох чи більше аргументів ($y = f(k, z, v)$).

Залежність функції від аргументу за лінійної регресії виражається коефіцієнтом регресії (b_{yx} , або b_{xy}), який показує, в якому напрямі і на яку величину змінюється одна ознака за зміни іншої на одиницю виміру:

$$b_{yx}b_{xy} = r^2 = d_{yx}$$

Завдання. Визначіть коефіцієнт прямолінійної кореляції та складіть рівняння регресії залежності вмісту цукру (у) від розміру коренеплоду цукрових буряків (х). Зробіть висновок і графічну інтерпретацію регресивної залежності (табл. 8.1).

1. Перший крок аналізу полягає у розрахунку допоміжних величин, похідних від первинних експериментальних даних.

Вміст цукру і маса коренеплоду буряків цукрових

№ пари величин	Величина ознак		x^2	y^2	xy
	x, g	$y, \%$			
1	157	16,2	24649	262,44	2543,4
2	245	17,4	60025	302,76	4263
3	543	14,7	294849	216,09	7982,1
4	347	17,8	120409	316,84	6176,6
5	379	17,7	143641	313,29	6708,3
6	410	16,5	168100	272,25	6765
7	720	14,2	518400	201,64	10224
8	215	17,6	46225	309,76	3784
9	310	18,0	96100	324	5580
10	640	14,5	409600	210,25	9280
Суми	3966	164,6	1881998	2729,32	63306,4

2. Визначення основних критеріїв кореляційного зв'язку

Кількість порівнюваних варіантів: $n = 10$. Середня величина x :
 $\bar{x} = \sum x : n = 3966 : 10 = 396,6$ Середня величина y : $\bar{y} = \sum y : n = 164,6 : 10 = 16,5$

Суми квадратів відхилень x і y від їхнього середнього показника:

$$\sum (x - \bar{x})^2 = \sum x^2 - (\sum x)^2 : n = 1881998 - (3966)^2 : 10 = 309082,4$$

$$\sum (y - \bar{y})^2 = \sum y^2 - (\sum y)^2 : n = 2729,32 - (164,6)^2 : 10 = 20,004$$

Добуток відхилень від їхнього середнього показника

$$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - (\sum x \sum y) : n = 63306,4 - (3966 \times 164,6) : 10 = -1973,96.$$

Коефіцієнт кореляції:

$$r = \frac{\sum xy - \sum x \sum y / n}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{-1973,96}{\sqrt{309082,4 \times 20,004}} = -0,79$$

3. Розрахунок коефіцієнта регресії та рівняння регресійної залежності:

Коефіцієнт регресії:

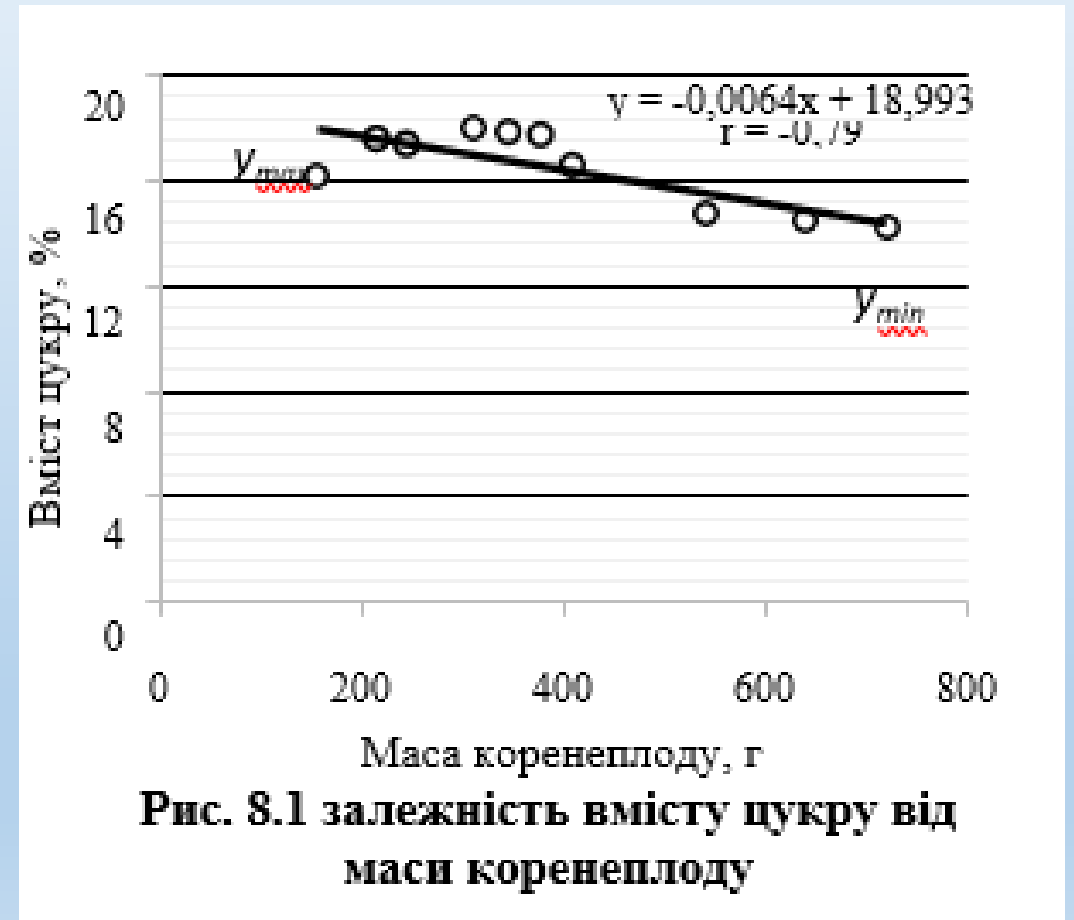
$$b_{yx} = \frac{\sum xy - \sum x \sum y / n}{\sum (x - \bar{x})^2} = \frac{-1973,96}{309082,4} = -0,006 \text{ \%}/2$$

Рівняння регресії і визначення величин y_{min} та y_{max} для побудови теоретичної лінії регресії:

$$y = \bar{y} + b_{xy}(x - \bar{x}) = 16,5 + (-0,006) * (x - 396,6) = 18,99 - 0,006x$$

$$y_{min} = 18,99 - 0,006 \times 640 = 15,1 \text{ \%}$$

$$y_{max} = 18,99 - 0,006 \times 157 = 18,0 \text{ \%}$$



Виявлення істотності кореляції та довірчих інтервалів

Помилка коефіцієнта кореляції:

$$S_r = \sqrt{(1 - r^2)/(n - 2)} = \sqrt{(1 - 0,79^2)/(10 - 2)} = 0,21 \% / 2$$

Помилка коефіцієнта регресії:

$$S_{b_{yx}} = S_r \sqrt{\sum (y - \bar{y})^2 / \sum (x - \bar{x})^2} = 0,21 * \sqrt{20,004 / 309082,4} = 0,0017 \% / 2$$

Відхилення лінії регресії:

$$S_{yx} = S_r \sqrt{\sum (y - \bar{y})^2} = 0,21 * \sqrt{20,004} = 0,96 \%$$

Критерій істотності коефіцієнта кореляції:

$$t_r = r / S_r = 0,79 / 0,21 = 3,69 \text{ при } \gamma = n - 2 = 10 - 2 = 8, t_{05} = 2,31$$

Довірчі інтервали коефіцієнтів:

$$\text{кореляції: } r \pm t_{05} * S_r = -0,79 \pm 2,31 * 0,21 = -0,79 \pm 0,48 = -0,31 \div -1$$

$$\begin{aligned} \text{регресії: } b_{yx} \pm t_{05} * S_{b_{yx}} &= -0,006 \pm 2,31 * 0,0017 = -0,006 \pm 0,004 = \\ &= -0,002 \div -0,01 \% / 2 \end{aligned}$$

Висновок. Між вмістом цукру та масою коренеплодів буряків цукрових виявлений тісний, зворотній, істотний кореляційний зв'язок на 5 %-му рівні значущості ($t_r > t_{05}$). Довірчі інтервали коефіцієнтів кореляції та регресії на 5 %-му рівні значущості свідчать про перебування величини критерію кореляції в інтервалі між середньою і сильною.