Т е м а 6. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

Объяснение. В биологических исследованиях редко приходится иметь дело с определенными функциональными связями между признаками, когда каждому значению одной величины соответствует строго определенное значение другой величины. Чаще всего каждому значению признака *X* соответствует не одно, а множество возможных значений признака *Y*. Такие связи называются *стохастическими* (вероятностными) или корреляционными.

Термин «корреляция» происходит от английского слова *correlation*, что означает соотношение, соответствие. Понятие корреляции введено в науку английским учёным Ф. Гальтоном (1888 г.) и развито его учеником К. Пирсоном (1895 г.). К изучению связи методом корреляции обращаются в том случае, когда невозможно элиминировать (изолировать) влияние посторонних факторов либо потому, что они неизвестны, либо из-за невозможности их изоляции. Кроме того метод позволяет найти и оценить количественную меру тесноты связи между взаимосвязанными признаками. При этом численность выборки должна быть достаточно большой, так как малое количество наблюдений не позволяет обнаружить закономерность связи.

Корреляция бывает простой (зависимость между двумя признаками) и множественной (больше двух), по форме – прямолинейной и криволинейной, по направлению – прямой и обратной.

Под прямолинейной корреляцией понимают такое соотношение между переменными, которое выражается уравнением прямой линии
$Y = a + bX$. Когда при одинаковых приращениях аргумента функция имеет неодинаковые изменения, корреляция называется криволинейной. Если при увеличении аргумента функция возрастает, то корреляция называется положительной или прямой, а если убывает – отрицательной или обратной.

В качестве числового показателя простой линейной корреляции, показывающего тесноту (силу) и направление связи, используют безразмерное число, называемое коэффициентом корреляции (*r*).

Значения коэффициента корреляции могут находиться в пределах от + 1 при прямой функциональной связи до - 1 при обратной функциональной связи. При полном отсутствии

корреляции $r = 0$, при $|r| < 0{,}3$ корреляционная зависимость слабая, при $|r| = 0{,}3–0{,}7$ – средняя, а при $|r| > 0{,}7$ – сильная. Знак при коэффициенте корреляции указывает направление связи: (+) – прямая зависимость, (-) – связь обратная.

Для анализа линейной корреляции между X и Y проводят n независимых парных наблюдений, исходом каждого из которых является пара чисел (X_1; Y_1), (X_2; Y_2), … (X_n; Y_n). Способ вычисления коэффициента корреляции рассмотрим на примере (табл. 5).

Таблица 5 – Вычисление коэффициента корреляции между продуктивностью растений и числом зерен в колосе

№ растения	Значение признаков		Отклонение от средней		Квадраты отклонений		Произведения отклонений
	Продуктивность растений Y, г/раст.	Число зерен в колосе X, шт.	$(Y - \bar{y})$	$(X - \bar{x})$	$(Y - \bar{y})^2$	$(X - \bar{x})^2$	$(Y - \bar{y}) \times (X - \bar{x})$
1	1,74	38	-0,13	-3	0,0169	9	0,39
2	2,06	46	0,19	5	0,0361	25	0,25
3	1,75	38	-0,12	-3	0,0144	9	0,36
4	2,00	42	0,13	1	0,0169	1	0,13
5	1,53	38	-0,34	-3	0,1156	9	1,02
6	1,78	44	-0,09	3	0,0081	9	-0,27
7	1,77	38	-0,10	-3	0,0100	9	0,30
8	1,80	37	-0,07	-4	0,0049	16	0,28
9	2,22	48	0,35	7	0,1225	49	2,45
10	2,05	41	0,18	0	0,0324	0	0
$\Sigma =$	18,70	410	0	0	0,3778	136	5,61
	$\bar{y} = 1{,}87$	$\bar{x} = 41$	$n = 10$				

Коэффициент корреляции вычисляют по формуле:

$$r = \frac{\sum(X - \bar{x})(Y - \bar{y})}{\sqrt{\sum(X - \bar{x})^2 \sum(Y - \bar{y})^2}} = \frac{5{,}61}{\sqrt{0{,}3778 \times 136}} = \frac{5{,}61}{\sqrt{51{,}3008}} = 0{,}78$$

или, минуя вычисления отклонений и квадратов отклонений, по формуле:

$$r = \frac{\sum XY - (\sum X \times \sum Y) \times n}{\sqrt{(\sum X^2 - (\sum X)^2 \div n)(\sum Y^2 - (\sum Y)^2 \div n)}}$$

или через средние квадратические отклонения –

$$r = \frac{\sum(X - \bar{x})(Y - \bar{y})}{n \times S_{\bar{x}} \times S_{\bar{y}}}$$

где: $S_{\bar{x}} = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$, $S_y = \sqrt{\frac{\sum(Y-\bar{y})^2}{n-1}}$

где: $(X - \bar{x})$ и $(Y - \bar{y})$ – отклонения значений X и Y от своих средних значений $\bar{x}$ и $\bar{y}$ в n сопоставимых парах.

Таким образом, связь между продуктивностью растений озимой ржи сорта Калинка и числом зерен в их колосьях сильная прямая (r = 0,78).

Степень связи между признаками более точно измеряется коэффициентом детерминации d_{yx}, равным квадрату коэффициента корреляции: $d_{yx} = r^2$. Он показывает долю тех изменений (%), которые зависят от изучаемого фактора. В нашем примере d_{yx} = $0{,}78^2$= 0,61, следовательно только 61% изменчивости признака Y обусловлено действием факториального признака X (числом зерен в колосе), остальная часть корреляционной связи (1 – 0,61 = 0,39) обусловлена другими факторами.

Коэффициент корреляции выборочных наблюдений подвержен случайным колебаниям, которые зависят от объема выборки и точности проведения наблюдений. Поэтому для оценки надежности выборочного коэффициента вычисляют его ошибку и критерий существенности.

Стандартную ошибку коэффициента корреляции определяют по формуле:

$$S_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0{,}78^2}{10-2}} = \pm 0{,}22$$

где: S_r – ошибка коэффициента корреляции;

r – коэффициент корреляции;

n – число пар значений.

Чем больше число наблюдений, тем меньше будет ошибка коэффициента корреляции. Значение коэффициента корреляции обычно записывается вместе с его ошибкой:

$$r \pm S_r = 0{,}78 \pm 0{,}22.$$

Критерий существенности (оценку значимости) коэффициента корреляции вычисляют по формуле:

$$t_r = \frac{r}{S_r} = \frac{0{,}78}{0{,}22} = 3{,}55$$

Сопоставляя фактические и теоретически рассчитанные значения t_r при числе степеней свободы, равном $n - 2$, оценивают существенность корреляционной связи. Если $t_{r\ \text{факт.}} \geq t_{r\ \text{теор.}}$, то корреляционная связь существенна, а при $t_{r\ \text{факт.}} < t_{r\ \text{теор.}}$ – несущественна. Теоретическое значение критерия Стьюдента

находят по таблице (приложение 1), принимая 5%-ный или 1%-ный уровень значимости. В нашем примере при восьми степенях свободы ($n - 2 = 8$) t_{05} (2,31) $\leq t_{r\ факт}$. (3,55). Значит, коэффициент корреляции в нашем случае является статистически значимым.

Определив коэффициент корреляции, мы выясняем направление и степень сопряженности в изменчивости признаков. Однако, он не позволяет узнать, как количественно изменяется результативный признак при изменении факториального на единицу измерения. Это решается с помощью регрессионного анализа. Его основная задача - определить формулу корреляционной зависимости. Различают регрессию простую и множественную, а по форме – прямо- и криволинейную. Сущность регрессионного анализа состоит в том, чтобы построить линию, которая наиболее точно выражала бы зависимость одного признака от другого.

Зависимость между признаками может быть выражена коэффициентом регрессии, показывающим, в каком направлении и на какую величину изменяется в среднем один признак (функция) при изменении другого (аргумент) на единицу измерения. Коэффициенты регрессии имеют знак коэффициента корреляции и вычисляются по следующим формулам:

$$b_{yx} = \frac{\sum(X - \bar{x})(Y - \bar{y})}{\sum(X - \bar{x})^2} = \frac{5{,}61}{136} = 0{,}04125$$

$$b_{xy} = \frac{\sum(X - \bar{x})(Y - \bar{y})}{\sum(Y - \bar{y})^2} = \frac{5{,}61}{0{,}3778} = 14{,}85$$

Их можно вычислить и через среднеквадратичные отклонения:

$$b_{yx} = r\frac{S_{\bar{y}}}{S_{\bar{x}}}, \quad b_{xy} = r\frac{S_{\bar{x}}}{S_{\bar{y}}}$$

Произведение коэффициентов регрессии равно квадрату коэффициента корреляции:

$$b_{yx} \times b_{xy} = r^2$$

$$r = \sqrt{0{,}04125 \times 14{,}85} = \sqrt{0{,}61} = 0{,}78$$

Ошибку коэффициентов регрессии вычисляют по формулам:

$$S_{b_{yx}} = S_r \cdot \sqrt{\frac{\sum(Y - \bar{y})^2}{\sum(X - \bar{x})^2}} = 0{,}22 \cdot \sqrt{\frac{0{,}3778}{136}} = 0{,}22 \times 0{,}0527 = 0{,}0116$$

$$S_{b_{xy}} = S_r \cdot \sqrt{\frac{\sum(X - \bar{x})^2}{\sum(Y - \bar{y})^2}} = 0{,}22 \cdot \sqrt{\frac{136}{0{,}3778}} = 0{,}22 \times 18{,}97 = 4{,}174$$

Критерий существенности коэффициента регрессии определяется по формуле:

$$t_{b_{yx}} = \frac{b_{yx}}{S_{b_{yx}}} = \frac{0{,}04125}{0{,}0116} = 3{,}55$$

Критерий существенности коэффициентов регрессии равен критерию существенности коэффициента корреляции: $t_{byx} = t_r$.

В зависимости от того, между какими признаками рассматривается связь, не всегда имеет смысл вычислять все коэффициенты регрессии. Линию регрессии можно построить двумя способами – графическим и аналитическим.

При графическом способе по оси абсцисс откладывают значения признака *X*, по оси ординат – значения признака *Y*. Такой график называют «точечной диаграммой» или «корреляционным полем».

При аналитическом способе используют уравнения прямой линии (для линейной регрессии):

$$Y = a + bX; \quad a = y - b\bar{x}; \quad b = b_{yx}$$

Уравнение линейной регрессии имеет следующий вид:

$$Y = \bar{y} + b_{yx}(X - \bar{x})$$

$Y = 1{,}87 + 0{,}04125\,(X - \bar{x}) = 1{,}87 + 0{,}04125\,X - 0{,}04125 \times 41 = 0{,}18 + 0{,}04125X$.

Выводы. Корреляционная зависимость между продуктивностью растений озимой ржи сорта Калинка и числом зерен в колосе прямая, сильная. Продуктивность растений на 61% зависит от числа зерен. При увеличении числа зерен в колосе на 1 шт. продуктивность растений увеличивается на 0,04125 г. Данная зависимость выражается уравнением $Y = 0{,}18 + 0{,}04125\,X$. Далее следует выполнить индивидуальное **задание** (приложение 8).