

§ 2. Криволинейная корреляция

Если график регрессии изображается кривой линии, то корреляцию называют *криволинейной*. В частности, в случае параболической корреляции второго порядка *выборочное уравнение регрессии* Y на X имеет вид

$$\bar{y}_x = Ax^2 + Bx + C.$$

Неизвестные параметры A , B и C находят (например, методом Гаусса) из системы уравнений:

$$\begin{cases} (\sum n_x x^4) A + (\sum n_x x^3) B + (\sum n_x x^2) C = \sum n_x \bar{y}_x x^2, \\ (\sum n_x x^3) A + (\sum n_x x^2) B + (\sum n_x x) C = \sum n_x \bar{y}_x x, \\ (\sum n_x x^2) A + (\sum n_x x) B + nC = \sum n_x \bar{y}_x. \end{cases} (*)$$

Аналогично находится выборочное уравнение регрессии X на Y

$$\bar{x}_y = A_1 y^2 + B_1 y + C_1.$$

Для оценки силы корреляции Y на X служит *выборочное корреляционное отношение* (отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению признака Y)

$$\eta_{yx} = \frac{\sigma_{\text{межгр}}}{\sigma_{\text{обш}}},$$

или (в других обозначениях)

$$\eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y}.$$

Здесь

$$\sigma_{\bar{y}_x} = \sqrt{D_{\text{межгр}}} = \sqrt{\frac{\sum n_x (\bar{y}_x - \bar{y})^2}{n}}, \quad \sigma_y = \sqrt{D_{\text{общ}}} = \sqrt{\frac{\sum n_y (y - \bar{y})^2}{n}},$$

где n — объем выборки (сумма всех частот); n_x — частота значения x признака X ; n_y — частота значения y признака Y ; \bar{y}_x — условная средняя признака Y ; \bar{y} — общая средняя признака Y .

Аналогично определяется выборочное корреляционное отношение X к Y :

$$\eta_{xy} = \frac{\sigma_{\bar{x}_y}}{\sigma_x}.$$

500. Найти выборочное уравнение регрессии $\bar{y}_x = Ax^2 + Bx + C$ по данным, приведенным в корреляционной таблице 8.

Оценить силу корреляционной связи по выборочному корреляционному соотношению.

	X	2	3	5	n_y
Y					20
25		20	—	—	31
45		—	30	1	49
110		—	31	48	
n_x		20	31	49	$n = 100$

Решение. Составим расчетную таблицу 9.

Таблица 9

\bar{x}	n_x	\bar{y}_x	$n_x x$	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \bar{y}_x$	$n_x \bar{y}_x x$	$n_x \bar{y}_x x^2$
2	20	25	40	80	160	320	500	1 000	2 000
3	31	47,1	93	279	837	2 511	4380	4 380	13 141
5	49	108,67	245	1225	6125	30 625	5325	26 624	133 121
Σ	100		378	1584	7122	33 456	7285	32 004	148 262

Подставив числа, содержащиеся в последней строке табл. 9, в (*), получим систему уравнений относительно неизвестных коэффициентов A , B , C :

$$\begin{aligned} 33456 A + 7122 B + 1584 C &= 148262, \\ 7122 A + 1584 B + 378 C &= 32004, \\ 1584 A + 378 B + 100 C &= 7285. \end{aligned}$$

Решив эту систему (например, методом Гаусса), найдем

$$A = 2,94, \quad B = 7,27, \quad C = -1,25.$$

Подставив найденные коэффициенты в уравнение регрессии

$$\bar{y}_x = Ax^2 + Bx + C,$$

окончательно получим

$$\bar{y}_x = 2,94x^2 + 7,27x - 1,25.$$

21.01.2014 11:45

§ 2. Криволинейная корреляция

Если график регрессии изображается кривой линии, то корреляцию называют *криволинейной*. В частности, в случае параболической корреляции второго порядка *выборочное уравнение регрессии* Y на X имеет вид

$$\bar{y}_x = Ax^2 + Bx + C.$$

Неизвестные параметры A , B и C находят (например, методом Гаусса) из системы уравнений:

$$\begin{cases} (\sum n_x x^4) A + (\sum n_x x^3) B + (\sum n_x x^2) C = \sum n_x \bar{y}_x x^2, \\ (\sum n_x x^3) A + (\sum n_x x^2) B + (\sum n_x x) C = \sum n_x \bar{y}_x x, \\ (\sum n_x x^2) A + (\sum n_x x) B + nC = \sum n_x \bar{y}_x. \end{cases} \quad (*)$$

Аналогично находится *выборочное уравнение регрессии* X на Y

$$\bar{x}_y = A_1 y^2 + B_1 y + C_1.$$

Для оценки силы корреляции Y на X служит *выборочное корреляционное отношение* (отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению признака Y)

$$\eta_{yx} = \frac{\sigma_{\text{межгр}}}{\sigma_{\text{общ}}},$$

или (в других обозначениях)

$$\eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y}.$$

Здесь

$$\sigma_{\bar{y}_x} = \sqrt{D_{\text{межгр}}} = \sqrt{\frac{\sum n_x (\bar{y}_x - \bar{y})^2}{n}}, \quad \sigma_y = \sqrt{D_{\text{общ}}} = \sqrt{\frac{\sum n_y (y - \bar{y})^2}{n}},$$

где n — объем выборки (сумма всех частот); n_x — частота значения x признака X ; n_y — частота значения y признака Y ; \bar{y}_x — условная средняя признака Y ; \bar{y} — общая средняя признака Y .

Аналогично определяется *выборочное корреляционное отношение* X к Y :

$$\eta_{xy} = \frac{\sigma_{\bar{x}_y}}{\sigma_x}.$$

500. Найти *выборочное уравнение регрессии* $\bar{y}_x = Ax^2 + Bx + C$ по данным, приведенным в корреляционной таблице 8.

Оценить силу корреляционной связи по *выборочному корреляционному соотношению*.

Таблица 8

	x	2	3	5	n_y
y					
		20	—	—	20
25		—	—	—	31
45		—	30	1	49
110		—	31	48	
n_x		20	31	49	$n = 100$

Решение. Составим расчетную таблицу 9.

Таблица 9

\bar{x}	n_x	\bar{y}_x	$n_x x$	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \bar{y}_x$	$n_x \bar{y}_x x$	$n_x \bar{y}_x x^2$
2	20	25	40	80	160	320	500	1 000	2 000
3	31	47,1	93	279	837	2 511	4380	4 380	13 141
5	49	108,67	245	1225	6125	30 625	5325	26 624	133 121
Σ	100		378	1584	7122	33 456	7285	32 004	148 262

Подставив числа, содержащиеся в последней строке табл. 9, в (*), получим систему уравнений относительно неизвестных коэффициентов A , B , C :

$$\begin{aligned} 33456 A + 7122 B + 1584 C &= 148262, \\ 7122 A + 1584 B + 378 C &= 32004, \\ 1584 A + 378 B + 100 C &= 7285. \end{aligned}$$

Решив эту систему (например, методом Гаусса), найдем

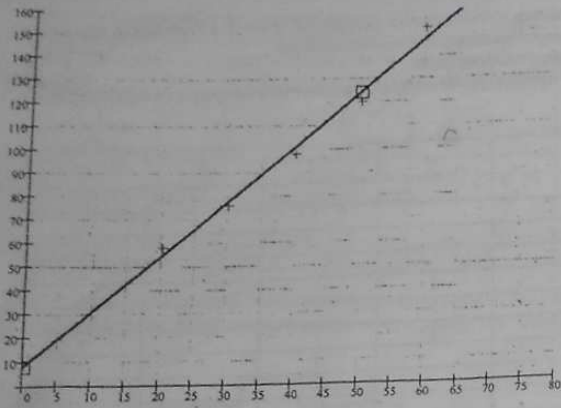
$$A = 2,94, \quad B = 7,27, \quad C = -1,25.$$

Подставив найденные коэффициенты в уравнение регрессии

$$\bar{y}_x = Ax^2 + Bx + C,$$

окончательно получим

$$\bar{y}_x = 2,94x^2 + 7,27x - 1,25.$$



Чтобы построить график прямой линейной регрессии найдём две точки лежащие на этой прямой: $\bar{Y}_0 = 8$, $\bar{Y}_{30} = 124$. Отмечаем условные средние (+) и эту прямую на координатной плоскости. Как видно из чертежа прямая проходит достаточно близко от условных средних, причём условные средние находятся по обе стороны от прямой, значит расчёты выполнены достаточно точно.

б) Найдем выборочный коэффициент корреляции

$$r_o = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\sigma_x \cdot \sigma_y} = \frac{4496 - 40 \cdot 100,8}{14,14214 \cdot 33,9317} = 0,96694.$$

Так как выборочный коэффициент корреляции близок к 1, то связь между высотой и массой данного растения достаточно тесная, близкая к функциональной линейной.

в) Найдем наблюдаемое значение критерия

$$t_{\text{набл}} = r_o \cdot \sqrt{\frac{n-2}{1-r_o^2}} = 0,96694 \cdot \sqrt{\frac{25-2}{1-0,96694^2}} = 18,18513.$$

По приложению 5 при заданном уровне значимости $p=0,05$ и $f=n-2=25-2=23$ степеней свободы находим соответствующее критическое значение $t_{\text{крит}}(p, f) = t_{\text{крит}}(0,05, 23) = 2,07$.

Т.к. $|t_{\text{набл}}| > t_{\text{крит}}$ то следует сделать вывод о значимости выборочного коэффициента корреляции (т.е. истинный коэффициент линейной корреляции существенно отличается от нуля).

Ответ: а) $\bar{Y}_x = 2,32 \cdot x + 8$;

б) 0,96694;

в) коэффициент корреляции значим.

Задание 11. Проверить при уровне значимости $p=0,05$ методами дисперсионного анализа эффективность воздействия количества вакцины (в дозах) на уровень заболеваемости (в %) по статистическим данным, приведенным в таблице.

Решение. У нас количество уровней $l=3$, а количество испытаний $q=5$. Для вычисления факторной дисперсии и остаточную дисперсии вычислим промежуточные данные.

Во-первых, рассчитаем групповые средние, о существенности различия которых мы хотим сделать выводы (см. по-

Номер испыт.	Количество вакцин (в условных ед.)		
	0	1	4
1	91	74	70
2	80	50	61
3	95	90	55
4	88	85	76
5	75	94	78
Ср. зн.	85,8	78,6	68

следнюю строку таблицы).

Во-вторых, найдем коэффициенты R_j и P_j .

$$R_j = \sum_{i=1}^q x_{ij} \cdot P_j = \sum_{i=1}^q x_{ij}^2.$$

$$R_1 = 91+80+95+88+75 = 429; \quad P_1 = 91^2+80^2+95^2+88^2+75^2 = 37075.$$

$$R_2 = 74+50+90+85+94 = 393; \quad P_2 = 74^2+50^2+90^2+85^2+94^2 = 32137.$$

$$R_3 = 70+61+55+76+78 = 340; \quad P_3 = 70^2+61^2+55^2+76^2+78^2 = 23506.$$

Найдём факторную дисперсию

$$s_{\text{факт}}^2 = \frac{1}{q(l-1)} \left(\sum_{j=1}^l R_j^2 - \frac{1}{l} \left(\sum_{j=1}^l R_j \right)^2 \right) = \frac{1}{5 \cdot 2} \left((429^2 + 393^2 + 340^2) - \frac{1}{3} (429 + 393 + 340)^2 \right) \approx 401.$$

Найдем остаточную дисперсию

$$s_{\text{ост}}^2 = \frac{1}{l(q-1)} \left(\sum_{j=1}^l P_j - \frac{1}{q} \sum_{j=1}^l R_j^2 \right) = \frac{1}{3 \cdot 4} \left((37075 + 32137 + 23506) - \frac{1}{5} (429^2 + 393^2 + 340^2) \right) \approx 158.$$

Далее воспользуемся правилом:

а) если $s_{\text{факт}}^2 < s_{\text{ост}}^2$ то следует сразу сделать вывод об отсутствии существенного влияния данных уровней фактора (количества вакцины) на процент заболевших людей.

б) в противном случае следует проверить значимость различия этих дисперсий при заданном уровне значимости и конкурирующей гипотезе $s_{\text{факт}}^2 > s_{\text{ост}}^2$. Т.е. по таблице распределения Фишера-Снедекора (прил.6) определяем Критическое значение критерия

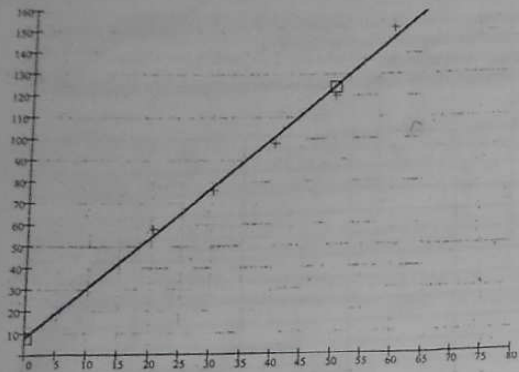
$$F_{\text{крит}} = F_{\text{крит}}(p, f_1=l-1, f_2=l(q-1)) = F_{\text{крит}}(0,05; 2; 12) = 3,88.$$

Сравниваем его с наблюдаемым значением критерия

$$F_{\text{набл}} = \frac{s_{\text{факт}}^2}{s_{\text{ост}}^2} = 401/158 = 2,5.$$

Т.к. $F_{\text{набл}} < F_{\text{крит}}$ то при данном уровне значимости принимаем гипотезу о равенстве $s_{\text{факт}}^2 = s_{\text{ост}}^2$, а потому следует сделать вывод о несущественности влияния данной вакцины на уровень заболеваемости (в случае $F_{\text{набл}} > F_{\text{крит}}$ делаем обратные выводы).

Ответ: влияние несущественно (т.е. очень вероятно, что уменьшение средних значений объясняется только случайностью выборки).



Чтобы построить график прямой линейной регрессии найдём две точки лежащие на этой прямой: $\bar{Y}_0=8$, $\bar{Y}_{50}=124$. Отмечаем условные средние (+) и эту прямую на координатной плоскости. Как видно из чертежа прямая проходит достаточно близко от условных средних, причём условные средние находятся по обе стороны от прямой, значит расчёты выполнены достаточно точно.

б) Найдем выборочный коэффициент корреляции

$$r_s = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\sigma_x \cdot \sigma_y} = \frac{4496 - 40 \cdot 100,8}{14,14214 \cdot 33,9317} = 0,96694.$$

Так как выборочный коэффициент корреляции близок к 1, то связь между высотой и массой данного растения достаточно тесная, близкая к функциональной линейной.

в) Найдем наблюдаемое значение критерия

$$t_{\text{набл}} = r_s \cdot \sqrt{\frac{n-2}{1-r_s^2}} = 0,96694 \cdot \sqrt{\frac{25-2}{1-0,96694^2}} = 18,18513.$$

По приложению 5 при заданном уровне значимости $p=0,05$ и $f=n-2=25-2=23$ степенях свободы находим соответствующее критическое значение $t_{\text{крит}}(p, f) = t_{\text{крит}}(0,05, 23) = 2,07$.

Т.к. $|t_{\text{набл}}| > t_{\text{крит}}$ то следует сделать вывод о значимости выборочного коэффициента корреляции (т.е. истинный коэффициент линейной корреляции существенно отличается от нуля).

Ответ: а) $\bar{Y}_x = 2,32x + 8$;

б) 0,96694;

в) коэффициент корреляции значим.

Задание 11. Проверить при уровне значимости $p=0,05$ методами дисперсионного анализа эффективность воздействия количества вакцины (в дозах) на уровень заболеваемости (в %) по статистическим данным, приведенным в таблице.

Решение. У нас количество уровней $l=3$, а количество испытаний $q=5$. Для вычисления факторной дисперсии и остаточную дисперсию вычислим промежуточные данные.

Во-первых, рассчитаем групповые средние, о существенности различия которых мы хотим сделать выводы (см. по-

Номер испыт.	Количество вакцин (в условных ед.)		
	0	1	4
1	91	74	70
2	80	50	61
3	95	90	55
4	88	85	76
5	75	94	78
Ср.зн.	85,8	78,6	68

следнюю строку таблицы).

Во-вторых, найдем коэффициенты R_j и P_j .

$$R_j = \sum_{i=1}^l x_{ij}, \quad P_j = \sum_{i=1}^l x_{ij}^2.$$

$$R_1 = 91+80+95+88+75 = 429; \quad P_1 = 91^2+80^2+95^2+88^2+75^2 = 37075.$$

$$R_2 = 74+50+90+85+94 = 393; \quad P_2 = 74^2+50^2+90^2+85^2+94^2 = 32137.$$

$$R_3 = 70+61+55+76+78 = 340; \quad P_3 = 70^2+61^2+55^2+76^2+78^2 = 23506.$$

Найдём факторную дисперсию

$$s_{\text{факт}}^2 = \frac{1}{q(l-1)} \left(\sum_{j=1}^l R_j^2 - \frac{1}{l} \left(\sum_{j=1}^l R_j \right)^2 \right) = \frac{1}{5 \cdot 2} \left((429^2 + 393^2 + 340^2) - \frac{1}{3} (429 + 393 + 340)^2 \right) = 401.$$

Найдём остаточную дисперсию

$$s_{\text{ост}}^2 = \frac{1}{l(q-1)} \left(\sum_{j=1}^l P_j - \frac{1}{q} \sum_{j=1}^l R_j^2 \right) = \frac{1}{3 \cdot 4} \left((37075 + 32137 + 23506) - \frac{1}{5} (429^2 + 393^2 + 340^2) \right) = 158.$$

Далее воспользуемся правилом:

а) если $s_{\text{факт}}^2 < s_{\text{ост}}^2$ то следует сразу сделать вывод об отсутствии существенного влияния данных уровней фактора (количества вакцины) на процент заболевших людей.

б) в противном случае следует проверить значимость различия этих дисперсий при заданном уровне значимости и конкурирующей гипотезе $s_{\text{факт}}^2 > s_{\text{ост}}^2$. Т.е. по таблице распределения Фишера-Снедекора (прил.6) определяем Критическое значение критерия

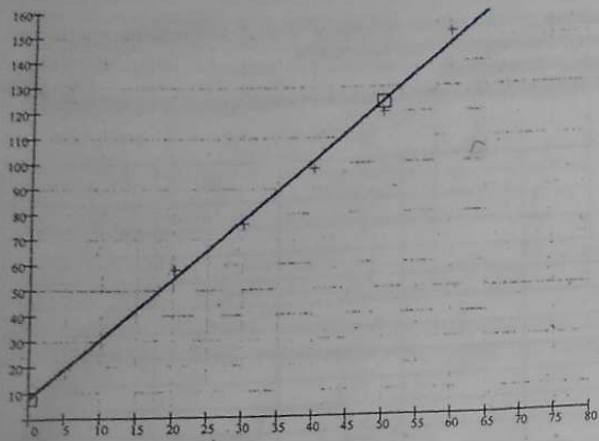
$$F_{\text{крит}} = F_{\text{крит}}(p, f_1=l-1, f_2=l(q-1)) = F_{\text{крит}}(0,05; 2; 12) = 3,88.$$

Сравниваем его с наблюдаемым значением критерия

$$F_{\text{набл}} = \frac{s_{\text{факт}}^2}{s_{\text{ост}}^2} = 401/158 = 2,5.$$

Т.к. $F_{\text{набл}} < F_{\text{крит}}$ то при данном уровне значимости принимаем гипотезу о равенстве $s_{\text{факт}}^2 = s_{\text{ост}}^2$, а потому следует сделать вывод о несущественности влияния данной вакцины на уровень заболеваемости (в случае $F_{\text{набл}} > F_{\text{крит}}$ делаем обратные выводы).

Ответ: влияние несущественно (т.е. очень вероятно, что уменьшение средних значений объясняется только случайностью выборки).



б) Найдем выборочный коэффициент корреляции

$$r_x = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\sigma_x \cdot \sigma_y} = \frac{4496 - 40 \cdot 100,8}{14,14214 \cdot 33,9317} = 0,96694.$$

Так как выборочный коэффициент корреляции близок к 1, то связь между высотой и массой данного растения достаточно тесная, близкая к функциональной линейной.

в) Найдем наблюдаемое значение критерия

$$t_{\text{набл}} = r_x \cdot \sqrt{\frac{n-2}{1-r_x^2}} = 0,96694 \cdot \sqrt{\frac{25-2}{1-0,96694^2}} = 18,18513.$$

По приложению 5 при заданном уровне значимости $p=0,05$ и $f=n-2=25-2=23$ степенях свободы находим соответствующее критическое значение $t_{\text{крит}}(p,f) = t_{\text{крит}}(0,05;23) = 2,07$.

Т.к. $|t_{\text{набл}}| > t_{\text{крит}}$ то следует сделать вывод о значимости выборочного коэффициента корреляции (т.е. истинный коэффициент линейной корреляции существенно отличается от нуля).

Ответ: а) $\bar{Y}_x = 2,32 \cdot x + 8$;

б) 0,96694;

в) коэффициент корреляции значим.

Задание 11. Проверить при уровне значимости $p=0,05$ методами дисперсионного анализа эффективность воздействия количества вакцины (в дозах) на уровень заболеваемости (в %) по статистическим данным, приведенным в таблице.

Решение. У нас количество уровней $l=3$, а количество испытаний $q=5$. Для вычисления факторной дисперсии и остаточную дисперсии вычислим промежуточные данные.

Во-первых, рассчитаем групповые средние, о существенности различия которых мы хотим сделать выводы (см. по

Номер испыт.	Количество вакцины (в условных ед.)		
	0	1	4
1	91	74	70
2	80	50	61
3	95	90	55
4	88	85	76
5	75	94	78
Ср.зн.	85,8	78,6	68

Чтобы построить график прямой линейной регрессии найдём две точки лежащие на этой прямой: $\bar{Y}_0 = 8$, $\bar{Y}_{50} = 124$. Отмечаем условные средние (+) и эту прямую на координатной плоскости. Как видно из чертежа прямая проходит достаточно близко от условных средних, причём условные средние находятся по обе стороны от прямой, значит расчёты выполнены достаточно точно.

Во-вторых, найдем коэффициенты R_j и P_j .

$$R_j = \sum_{i=1}^q x_{ij}, P_j = \sum_{i=1}^q x_{ij}^2.$$

$$R_1 = 91+80+95+88+75 = 429; P_1 = 91^2+80^2+95^2+88^2+75^2 = 37075.$$

$$R_2 = 74+50+90+85+94 = 393; P_2 = 74^2+50^2+90^2+85^2+94^2 = 32137.$$

$$R_3 = 70+61+55+76+78 = 340; P_3 = 70^2+61^2+55^2+76^2+78^2 = 23506.$$

Найдём факторную дисперсию

$$s_{\text{факт}}^2 = \frac{1}{q(l-1)} \left(\sum_{j=1}^l R_j^2 - \frac{1}{l} \left(\sum_{j=1}^l R_j \right)^2 \right) = \frac{1}{5 \cdot 2} \left((429^2 + 393^2 + 340^2) - \frac{1}{3} (429 + 393 + 340)^2 \right) \approx 401.$$

Найдём остаточную дисперсию

$$s_{\text{ост}}^2 = \frac{1}{l(q-1)} \left(\sum_{j=1}^l P_j - \frac{1}{q} \sum_{j=1}^l R_j^2 \right) = \frac{1}{3 \cdot 4} \left((37075 + 32137 + 23506) - \frac{1}{5} (429^2 + 393^2 + 340^2) \right) \approx 158.$$

Далее воспользуемся правилом:

а) если $s_{\text{факт}}^2 < s_{\text{ост}}^2$ то следует сразу сделать вывод об отсутствии существенного влияния данных уровней фактора (количества вакцины) на процент заболевших людей.

б) в противном случае следует проверить значимость различия этих дисперсий при заданном уровне значимости и конкурирующей гипотезе $s_{\text{факт}}^2 > s_{\text{ост}}^2$. Т.е. по таблице распределения Фишера-Снедекора (прил.6) определяем Критическое значение критерия

$$F_{\text{крит}} = F_{\text{крит}}(p, f_1=l-1, f_2=l(q-1)) = F_{\text{крит}}(0,05;2;12) = 3,88.$$

Сравниваем его с наблюдаемым значением критерия

$$F_{\text{набл}} = \frac{s_{\text{факт}}^2}{s_{\text{ост}}^2} = 401/158 = 2,5.$$

Т.к. $F_{\text{набл}} < F_{\text{крит}}$ то при данном уровне значимости принимаем гипотезу о равенстве $s_{\text{факт}}^2 = s_{\text{ост}}^2$, а потому следует сделать вывод о несущественности влияния данной вакцины на уровень заболеваемости (в случае $F_{\text{набл}} > F_{\text{крит}}$ делаем обратные выводы).

Ответ: влияние несущественно (т.е. очень вероятно, что уменьшение средних значений объясняется только случайностью выборки).

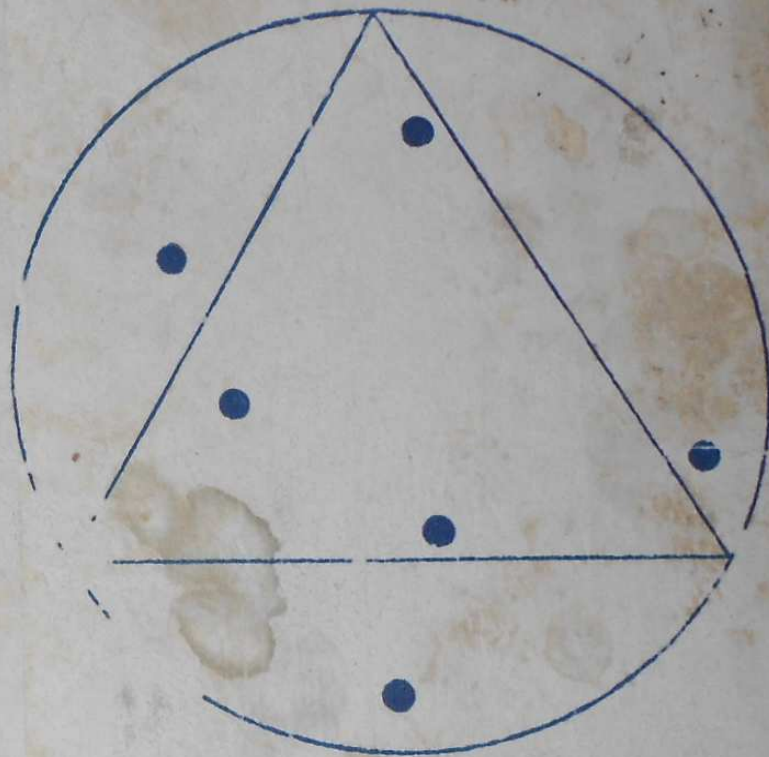
21.01.2014 11:46

53 коп.

ИЗДАТЕЛЬСТВО МОС КВА
1975 ГОД
ВЫСШАЯ ШКОЛА



В. Е. ГМУРМАН



В. Е. ГМУРМАН

РУКОВОДСТВО
К РЕШЕНИЮ
ЗАДАЧ
ПО ТЕОРИИ
ВЕРОЯТНОСТЕЙ
И МАТЕМАТИЧЕСКОЙ
СТАТИСТИКЕ

21.01.2014 11:46

Чтобы построить график регрессии, надо выбрать две точки, лежащие на этой прямой: $\bar{Y}_1 = 8$, $\bar{Y}_2 = 124$. Отметим условные средние (\bar{x}, \bar{y}) и эту прямую на координатной плоскости. Как видно из чертежа, прямая проходит достаточно близко от условных средних, причем условные средние находятся по обе стороны от прямой, значит расчеты выполнены достаточно точно.

б) Найдем выборочный коэффициент корреляции

$$r_{xy} = \frac{\bar{X}\bar{Y} - \bar{X}\bar{Y}}{\sigma_x \sigma_y} = \frac{4496 - 40 \cdot 100,8}{14,14214 \cdot 33,917} = 0,96694$$

Так как выборочный коэффициент корреляции близок к 1, то связь между высотой и массой данного растения достаточно тесная, близкая к функциональной линейной.

в) Найдем наблюдаемое значение критерия

$$t_{набл} = r_{xy} \sqrt{\frac{n-2}{1-r_{xy}^2}} = 0,96694 \sqrt{\frac{25-2}{1-0,96694^2}} = 18,18513$$

По приложению 2 при малом уровне значимости $r=0,05$ и $f_{кр} = 25 - 2 = 23$ степеней свободы мы находим соответствующее критическое значение $t_{кр}(0,05; 23) = 2,07$.

Т.к. $|t_{набл}| > t_{кр}$, то следует сделать вывод о значимости выборочного коэффициента корреляции (т.е. истинный коэффициент линейной корреляции существенно отличается от нуля).

Ответ а) $\bar{Y}_1 = 8,22 \pm 8$,
б) 0,96694,
в) коэффициент корреляции значим.

Задача 11. Проверить при уровне значимости $r=0,05$ методы дисперсионного анализа эффективности воздействия фактора А (в % от нормы) на урожайность (в %) по статистическим данным, приведенным в таблице.

Фактор А	Урожай, %			
	Номер опыта (в условных ед.)	Количество растений (в условных ед.)		
	0	1	4	4
1.	91	74	70	70
2.	80	50	61	61
3.	95	90	55	55
4.	88	85	76	76
5.	75	94	78	78
Ср.зн.	85,8	78,6	68	68

Решение. У нас количество уровней $k=5$, а количество испытаний $q=5$. Для выявления факторной дисперсии и остаточную дисперсию вычислим промежуточные данные.

Во-первых, рассчитаем групповые средние, о суммировании различия которых мы хотим убедиться.

следнюю строку таблицы).

Во-вторых, найдем коэффициенты R_i и P_j .

$$R_1 = 91 + 80 + 95 + 88 + 75 = 429; \quad P_1 = 91^2 + 80^2 + 95^2 + 88^2 + 75^2 = 37075, \\ R_2 = 74 + 50 + 90 + 85 + 94 = 393; \quad P_2 = 74^2 + 50^2 + 90^2 + 85^2 + 94^2 = 32137, \\ R_3 = 70 + 61 + 55 + 78 + 340 = 614; \quad P_3 = 70^2 + 61^2 + 55^2 + 78^2 + 340^2 = 23506.$$

Найдем факторную дисперсию

$$s_{факт}^2 = \frac{1}{q(q-1)} \left(\sum_{i=1}^k R_i^2 - \frac{1}{q} \left(\sum_{i=1}^k R_i \right)^2 \right) = \frac{1}{5 \cdot 4} \left(37075 + 32137 + 23506 - \frac{1}{5} (429^2 + 393^2 + 614^2) \right) = 158.$$

Найдем остаточную дисперсию

$$s_{ост}^2 = \frac{1}{(q-1) \left(\sum_{j=1}^q P_j - \frac{1}{q} \left(\sum_{j=1}^q P_j \right)^2 \right)} = \frac{1}{3 \cdot 4} \left(37075 + 32137 + 23506 - \frac{1}{3} (429^2 + 393^2 + 614^2) \right) = 401.$$

Далее воспользуемся правилом.

а) если $s_{факт}^2 < s_{ост}^2$, то следует сразу сделать вывод об отсутствии существенного влияния фактора на уровень фактора (количества растений) на процент заболевших людей.

б) в противном случае следует проверить значимость различия этих дисперсий при уровне значимости и конкурирующей гипотезе $s_{факт}^2 > s_{ост}^2$. Т.е. по таблице распределения Фишера-Снедекора (прил.б) определить критическое значение критерия $F_{крит} = F_{табл}(f_1=f_1, f_2=(q-1)) = F_{табл}(0,05; 2; 12) = 3,88$.

Сравниваем его с наблюдаемым значением критерия

$$F_{набл} = \frac{s_{факт}^2}{s_{ост}^2} = 401/158 = 2,5$$

Т.к. $F_{набл} < F_{крит}$, то при данном уровне значимости принимаем гипотезу о равенстве дисперсий $s_{факт}^2 = s_{ост}^2$, а потому следует сделать вывод о несущественности влияния данной фактора на уровень заболеваемости (в случае $F_{набл} > F_{крит}$ делаем обратные выводы).

Ответ: влияние несущественно (т.е. очень вероятно, что уменьшение средних значений объема является только случайностью выборки).

21.01.2014 11:51

§ 2. Криволинейная корреляция

Если график регрессии изображается кривой 2-го порядка, то корреляция называется криволинейной. В частности, в случае параболической корреляции второго порядка **выборочное уравнение регрессии** Y на X имеет вид

$$y_x = Ax^2 + Bx + C.$$

Неизвестные параметры A, B и C находят (например, методом Гаусса) из системы уравнений:

$$\begin{cases} (\sum n_i x_i^4) A + (\sum n_i x_i^3) B + (\sum n_i x_i^2) C = \sum n_i y_i x_i^2, \\ (\sum n_i x_i^3) A + (\sum n_i x_i^2) B + (\sum n_i x_i) C = \sum n_i y_i x_i, \\ (\sum n_i x_i^2) A + (\sum n_i x_i) B + nC = \sum n_i y_i. \end{cases} (*)$$

Аналогично находится выборочное уравнение регрессии X на Y : $x_y = A_1 y^2 + B_1 y + C_1$.

Для оценки силы корреляции Y на X служит **выборочное корреляционное отношение** (отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению признака Y)

$$\eta_{yx} = \frac{\sigma_{критр}}{\sigma_{общ}}$$

или (в других обозначениях)

$$\eta_{yx} = \frac{\sigma_y}{\sigma_y}$$

Здесь,

$$\sigma_{критр} = \sqrt{D_{критр}} = \sqrt{\frac{\sum n_i (y_i - \bar{y})^2}{n}}, \quad \sigma_{общ} = \sqrt{D_{общ}} = \sqrt{\frac{\sum n_i (y_i - \bar{y})^2}{n}}$$

т.е. n — объем выборки (сумма всех частот); n_i — частота значения x признака X ; n_{y_j} — частота значения y признака Y ; \bar{y}_x — условная средняя признака Y ; \bar{y} — общая средняя признака Y . Аналогично определяется выборочное корреляционное отношение X к Y :

$$\eta_{xy} = \frac{\sigma_{критр}}{\sigma_x}$$

500. Найти выборочное уравнение регрессии $y_x = Ax^2 + Bx + C$ по данным, приведенным в корреляционной таблице 8.

Оценить силу корреляционной связи по выборочному корреляционному отношению.

Таблица 8

$x \backslash y$	2	3	5	n_{y_j}
25	20	—	—	20
45	—	30	1	31
110	—	31	48	49
n_x	20	31	49	$n = 100$

Решение. Составим расчетную таблицу 9.

Таблица 9

i	n_x	\bar{y}_x	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \bar{y}_x$	$n_x \bar{y}_x x$	$n_x \bar{y}_x x^2$
2	20	25	40	80	160	500	1 000	2 000
3	31	47,1	93	279	837	2 511	4 380	13 141
5	49	108,67	245	1 225	6 125	5 325	26 624	133 121
Σ	100	—	378	1 584	7 122	33 456	72 85	148 262

Подставив числа, содержащиеся в последней строке табл. 9, в (*), получим систему уравнений относительно неизвестных коэффициентов A, B, C :

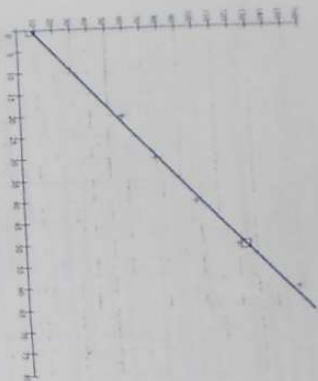
$$\begin{cases} 33456 A + 7122 B + 1584 C = 148262, \\ 7122 A + 1584 B + 378 C = 32004, \\ 1584 A + 378 B + 100 C = 7285. \end{cases}$$

Решив эту систему (например, методом Гаусса), найдем

$$A = 2,94, \quad B = 7,27, \quad C = -1,25.$$

Подставив найденные коэффициенты в уравнение регрессии окончательно получим

$$y_x = 2,94x^2 + 7,27x - 1,25.$$



6) Найти выборочный коэффициент корреляции

$$r_{xy} = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sigma_x \sigma_y} = \frac{4496 - 40 \cdot 100,8}{14,14214 \cdot 31,9317} = 0,96694$$

Так как выборочный коэффициент корреляции близок к 1, то связь между количеством и массой данного растения достаточно тесная, близкая к функциональной линейной.

ж) Найти выборочное значение критерия

$$t_{\text{выб}} = r_{xy} \sqrt{\frac{n-2}{1-r_{xy}^2}} = \frac{25-2}{1-0,96694^2} = -18,18513$$

По таблице 5 при данном уровне значимости $r_{\text{таб}}(0,05; n-2) = 0,25-0,23$ статистика больше, чем критическое значение $r_{\text{таб}}(0,05; 23) = 0,27$.
Т.к. $t_{\text{выб}} < r_{\text{таб}}$, то следует сделать вывод о значимости выборочного коэффициента корреляции (т.е. линейный коэффициент корреляции существенно отличается от нуля).

Ответ а) $r_{xy} = 0,96694$

б) коэффициент корреляции равен

Задание 11. Проверить при уровне значимости $p=0,05$ не только достоверность гипотезы эффективности, но и достоверность гипотезы равенства количества выходов (в среднем) на урожай Y_1 и Y_2 в % по статистическим данным, приведенным в таблице.

Культура	Число выходов (в %)	Число растений
1	91	74
2	80	50
3	95	50
4	88	85
5	75	94
Срм.	85,8	70,5

Вспомогательные группы вычислений, относящиеся к различным культурам, даны в таблице.

§ 2. Криволинейная корреляция

Если график регрессии изображается кривой дуги, то корреляцию называют криволинейной. В частности, в случае параболы второй степени криволинейная корреляция называется криволинейной корреляцией второго порядка.

$$g_{xy} = Ax^2 + Bx + C$$

Неизвестные параметры A, B и C находят (например, методом Гаусса) из системы уравнений:

$$\begin{cases} \sum (n_i x_i^2) A + \sum (n_i x_i^3) B + \sum (n_i x_i^4) C = \sum n_i d_i x_i^2 \\ \sum (n_i x_i^3) A + \sum (n_i x_i^4) B + \sum (n_i x_i^5) C = \sum n_i d_i x_i^3 \\ \sum (n_i x_i^4) A + \sum (n_i x_i^5) B + \sum (n_i x_i^6) C = \sum n_i d_i x_i^4 \end{cases} \quad (*)$$

Аналогично находят выборочное уравнение регрессии X на Y

$$x_{y_0} = Ay^2 + By + C$$

Для оценки связи корреляция Y на X служит *выборочное корреляционное отношение* (отношение между суммой квадратов вариации по отношению к общему среднему квадратическому отношению вариации Y)

$$\eta_{yx} = \frac{\sigma_{\text{крит}}}{\sigma_{\text{общ}}}$$

$$\eta_{yx} = \frac{\sigma_{\text{крит}}}{\sigma_y}$$

$$\sigma_{\text{крит}} = \sqrt{\frac{\sum n_i (d_i - \bar{d})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum n_i (y_i - \bar{y})^2}{n}}$$

где n — объем выборки (число всех случаев); n_i — частота значения x признака X ; n_{ij} — частота значения g признака Y ; d — условный признак Y ; \bar{d} — общий средний признак Y ; g_{xy} — условный признак Y . Аналогично определяется выборочное корреляционное отношение X к Y :

$$\eta_{xy} = \frac{\sigma_{\text{крит}}}{\sigma_x}$$

500. Найти выборочное уравнение регрессии $g_{xy} = Ax^2 + Bx + C$ по данным, приведенным в корреляционной таблице 8.

Оценить силу корреляционной связи по выборочному корреляционному отношению.

случайно строку таблицы)

Во-вторых, найти коэффициенты R и P

$$R = \sum_{i=1}^k x_i, P = \sum_{i=1}^k x_i^2$$

$$R = 91 + 80 + 95 + 88 + 75 = 429, P = 9^2 + 80^2 + 95^2 + 88^2 + 75^2 = 37075$$

$$R = 74 + 50 + 90 + 85 + 94 = 393; P = 74^2 + 50^2 + 90^2 + 85^2 + 94^2 = 32137$$

$$R = 70 + 61 + 55 + 76 + 78 + 40 = 328; P = 70^2 + 61^2 + 55^2 + 76^2 + 78^2 + 40^2 = 23506$$

Найдем факторное дисперсию

$$s_{\text{фак}}^2 = \frac{1}{q} \left(\sum_{i=1}^k R_i^2 - \frac{1}{q} \left(\sum_{i=1}^k R_i \right)^2 \right) = \frac{1}{5 \cdot 2} \left(429^2 + 393^2 + 340^2 - \frac{1}{3} (429^2 + 393^2 + 340^2)^2 \right)$$

Найдем остаточную дисперсию

$$s_{\text{ост}}^2 = \frac{1}{l(q-1)} \left(\sum_{i=1}^k P_i - \frac{1}{q} \sum_{i=1}^k R_i^2 \right) = \frac{1}{3 \cdot 2} \left(37075 + 32137 + 23506 - \frac{1}{3} (429^2 + 393^2 + 340^2)^2 \right)$$

Далее воспользуемся правилом:

а) если $s_{\text{фак}}^2 < s_{\text{ост}}^2$, то следует сразу сделать вывод об отсутствии существенного влияния фактора (количество выходов) на процент заболевших людей.

б) в противном случае следует проверить значимость различия этих дисперсий при помощи критерия Фишера-Снедекора (при этом $F_{\text{выб}} = \frac{s_{\text{фак}}^2}{s_{\text{ост}}^2} > F_{\text{таб}}$). Т.е. по таблице распределения функции Снедекора (приложение Критическое значение критерия Фишера-Снедекора (приложение К) $F_{\text{таб}} = F_{\text{выб}}(p; f_1-1; f_2-1) = F_{\text{таб}}(0,05; 2; 12) = 3,88$.

Сравним ее с таблическим значением критерия

$$F_{\text{выб}} = \frac{s_{\text{фак}}^2}{s_{\text{ост}}^2} = 401,158 > 2,5$$

Т.к. $F_{\text{выб}} > F_{\text{таб}}$, то при данном уровне значимости принимаем гипотезу о равенстве $s_{\text{фак}}^2 = s_{\text{ост}}^2$, а потому следует сделать вывод о несущественности влияния данной величины на уровень заболеваемости (в случае $F_{\text{выб}} > F_{\text{таб}}$ делаем обратные выводы).

Ответ: влияние несущественно (т.е. очень вероятно, что увеличение средних значений количества случаев случайностью заболеваемости).

Таблица 8

X	2	3	5	n_{xy}
25	20	—	—	20
45	—	30	1	31
110	—	31	48	49
n_x	20	31	49	$n = 100$

Решение. Составим расчетную таблицу 9.

i	n_{ix}	n_{ix}^2	n_{ix}^3	n_{ix}^4	n_{ix}^5	n_{ix}^6	n_{ix}^7
2	20	25	40	80	100	320	500
3	31	471	93	279	837	2 511	4 380
5	49	108 67	245	1225	6125	30 625	53 225
Σ	100	378	1384	7122	33 456	72 85	148 202

Поставив числа, содержащиеся в последней строке таблицы 9, в $(*)$, получим систему уравнений относительно коэффициентов A, B, C :

$$33456A + 7122B + 1384C = 148202$$

$$7122A + 1584B + 378C = 32004$$

$$1584A + 378B + 100C = 7285$$

Поставив систему (например, методом Гаусса), найдем

$$A = 2,94, B = 7,27, C = -1,25$$

окончательно найдем коэффициенты в уравнении регрессии

$$g_{xy} = 2,94x^2 + 7,27x - 1,25$$