§ 2. Криволинейная корреляция

Если график регрессии изображается кривой линией, 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:

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

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

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

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

Y \ X	2	3	5	n_y
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

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

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

§ 2. Криволинейная корреляция

Если график регрессии изображается кривой линией, 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 $\bar{y}$ 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

Y \ X	2	3	5	n_y
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

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

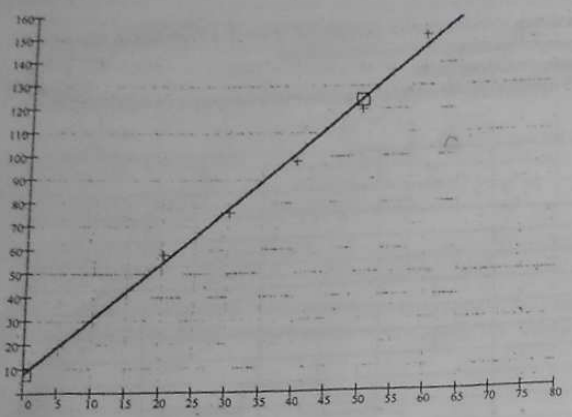
$$A = 2{,}94,\quad B = 7{,}27,\quad C = -1{,}25.$$

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 прямолинейной регрессии найдём две точки лежащие на этой прямой: $\overline{Y}_0=8$, $\overline{Y}_{50}=124$. 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

$$r_в=\frac{\overline{XY}-\overline{X}\cdot\overline{Y}}{\sigma_x\cdot\sigma_y}=\frac{4496-40\cdot100{,}8}{14{,}14214\cdot33{,}9317}=0{,}96694.$$

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

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

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

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

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

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

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. последнюю строку таблицы).

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

$$s^2_{факт}=\frac{1}{q(l-1)}\left(\sum_{j=1}^{l}R_j^2-\frac{1}{l}\left(\sum_{j=1}^{l}R_j\right)^2\right)=\frac{1}{5\cdot2}\left((429^2+393^2+340^2)-\frac{1}{3}(429+393+340)^2\right)\approx401.$$

Найдем остаточную дисперсию

$$s^2_{ост}=\frac{1}{l(q-1)}\left(\sum_{j=1}^{l}P_j-\frac{1}{q}\sum_{j=1}^{l}R_j^2\right)=\frac{1}{3\cdot4}\left((37075+32137+23506)-\frac{1}{5}(429^2+393^2+340^2)\right)\approx158.$$

Далее воспользуемся правилом:
а) если $s^2_{факт}<s^2_{ост}$ 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 $s^2_{факт}>s^2_{ост}$. Т.е. по таблице распределения Фишера-Снедекора (прил.6) определяем Критическое значение критерия

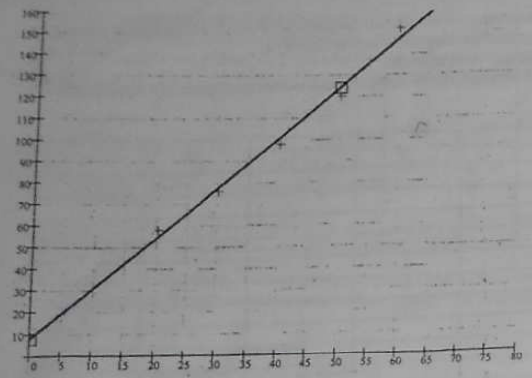
$$F_{крит}=F_{крит}(p,f_1=l-1,f_2=l(q-1))=F_{крит}(0{,}05;2;12)=3{,}88.$$

Сравниваем его с наблюдаемым значением критерия

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

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

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



Чтобы построить график прямолинейной регрессии найдём две точки лежащие на этой прямой: $\overline{Y}_0=8$, $\overline{Y}_{50}=124$. 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

$$r_в=\frac{\overline{XY}-\overline{X}\cdot\overline{Y}}{\sigma_x\cdot\sigma_y}=\frac{4496-40\cdot100,8}{14,14214\cdot33,9317}=0,96694.$$

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

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

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

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

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

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

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. последнюю строку таблицы).

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

$$s^2_{факт}=\frac{1}{q(l-1)}\left(\sum_{j=1}^{l}R_j^2-\frac{1}{l}\left(\sum_{j=1}^{l}R_j\right)^2\right)=\frac{1}{5\cdot2}\left((429^2+393^2+340^2)-\frac{1}{3}(429+393+340)^2\right)\approx401.$$

Найдем остаточную дисперсию

$$s^2_{ост}=\frac{1}{l(q-1)}\left(\sum_{j=1}^{l}P_j-\frac{1}{q}\sum_{j=1}^{l}R_j^2\right)=\frac{1}{3\cdot4}\left((37075+32137+23506)-\frac{1}{5}(429^2+393^2+340^2)\right)\approx158.$$

Далее воспользуемся правилом:
а) если $s^2_{факт}<s^2_{ост}$ 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 $s^2_{факт}>s^2_{ост}$. Т.е. по таблице распределения Фишера-Снедекора (прил.6) определяем Критическое значение критерия

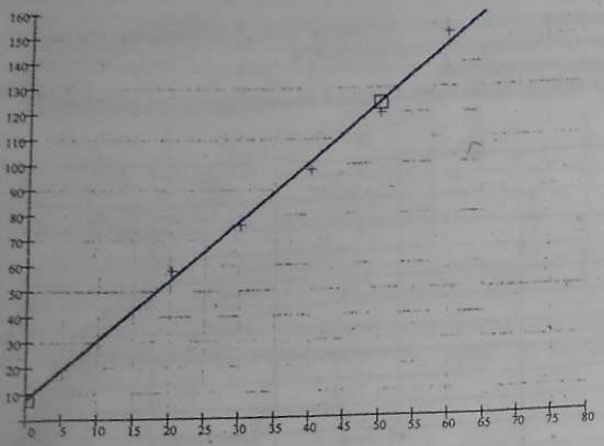
$$F_{крит}=F_{крит}(p,f_1=l-1,f_2=l(q-1))=F_{крит}(0,05;2;12)=3,88.$$

Сравниваем его с наблюдаемым значением критерия

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

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

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



Чтобы построить график прямолинейной регрессии найдём две точки лежащие на этой прямой: $\overline{Y}_0=8$, $\overline{Y}_{50}=124$. 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

$$r_в=\frac{\overline{XY}-\overline{X}\cdot\overline{Y}}{\sigma_x\cdot\sigma_y}=\frac{4496-40\cdot100,8}{14,14214\cdot33,9317}=0,96694.$$

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

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

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

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

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

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

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-

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

$$s^2_{факт}=\frac{1}{q(l-1)}\left(\sum_{j=1}^{l}R_j^2-\frac{1}{l}\left(\sum_{j=1}^{l}R_j\right)^2\right)=\frac{1}{5\cdot2}\left((429^2+393^2+340^2)-\frac{1}{3}(429+393+340)^2\right)\approx401.$$

Найдем остаточную дисперсию

$$s^2_{ост}=\frac{1}{l(q-1)}\left(\sum_{j=1}^{l}P_j-\frac{1}{q}\sum_{j=1}^{l}R_j^2\right)=\frac{1}{3\cdot4}\left((37075+32137+23506)-\frac{1}{5}(429^2+393^2+340^2)\right)\approx158.$$

Далее воспользуемся правилом:
а) если $s^2_{факт}<s^2_{ост}$ 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 $s^2_{факт}>s^2_{ост}$. Т.е. по таблице распределения Фишера-Снедекора (прил.6) определяем Критическое значение критерия

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

Сравниваем его с наблюдаемым значением критерия

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

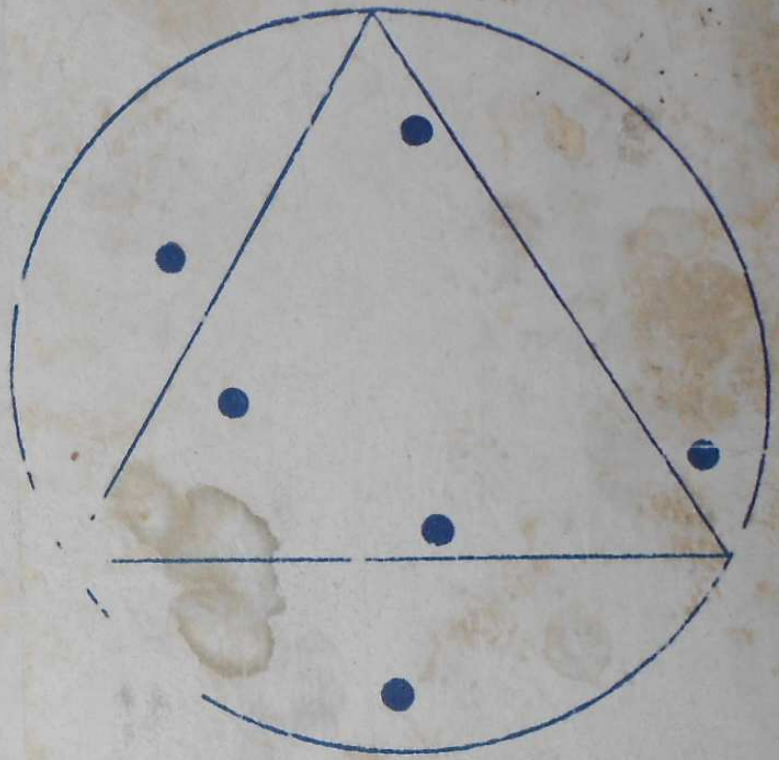
Т.к. $F_{набл}<F_{крит}$, то при данном уровне значимости принимаем гипотезу о равенстве $s^2_{факт}=s^2_{ост}$, а потому следует сделать вывод о несущественности влияния данной вакцины на уровень заболеваемости (в случае $F_{набл}>F_{крит}$ делаем обратные выводы).

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

53 коп.

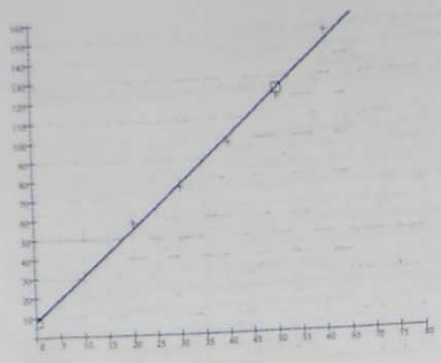
ИЗДАТЕЛЬСТВО МОСКВА
1975 ГОД ВЫСШАЯ ШКОЛА

В. Е. ГМУРМАН



В. Е. ГМУРМАН

РУКОВОДСТВО К РЕШЕНИЮ ЗАДАЧ ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ



Чтобы построить график прямолинейной регрессии найдём две точки лежащие на этой прямой: $\bar{Y}_0=8$, $\bar{Y}_{10}=124$. 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

$$r_в=\frac{\overline{XY}-\bar{X}\cdot\bar{Y}}{\sigma_x\cdot\sigma_y}=\frac{4496-40\cdot100{,}8}{14{,}14214\cdot33{,}9317}=0{,}96694.$$

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

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

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

Т.к. $|t_{набл}|>t_{крит}$ 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 возлействия количества вакцины (в дозах) на уровень заболеваемости (в %) по статистическим данным, приведенным в таблице.

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
Срзн.	85,8	78,6	68

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 вывод (см. последнюю строку таблицы).

Во-вторых, найдем коэффициенты R_j и P_j.

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

$R_1=91+80+95+88+75=429$; $P_1=91^2+80^2+95^2+88^2+75^2=37075$.
$R_2=74+50+90+85+94=393$; $P_2=74^2+50^2+90^2+85^2+94^2=32137$.
$R_3=70+61+55+76+78=340$; $P_3=70^2+61^2+55^2+76^2+78^2=23506$.

Найдём факторную дисперсию

$$s^2_{факт}=\frac{1}{q(l-1)}\left(\sum_{j=1}^{l}R_j^2-\frac{1}{l}\left(\sum_{j=1}^{l}R_j\right)^2\right)=\frac{1}{5\cdot2}\left((429^2+393^2+340^2)-\frac{1}{3}(429+393+340)^2\right)\approx401.$$

Найдем остаточную дисперсию

$$s^2_{ост}=\frac{1}{l(q-1)}\left(\sum_{j=1}^{l}P_j-\frac{1}{q}\sum_{j=1}^{l}R_j^2\right)=\frac{1}{3\cdot4}\left((37075+32137+23506)-\frac{1}{5}(429^2+393^2+340^2)\right)\approx158.$$

Далее воспользуемся правилом:
а) если $s^2_{факт}<s^2_{ост}$ 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 $s^2_{факт}>s^2_{ост}$. Т.е. по таблице распределения Фишера-Снедекора (прил.6) определяем Критическое значение критерия

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

Сравниваем его с наблюдаемым значением критерия

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

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

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

§ 2. Криволинейная корреляция

Если график регрессии изображается кривой линией, 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:

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

Аналогично находится выборочное уравнение регрессии X на Y

$$\bar{x}_y=A_1y^2+B_1y+C_1.$$

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

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

или (в других обозначениях)

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

Здесь

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

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

Y \ X	2	3	5	n_y
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

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

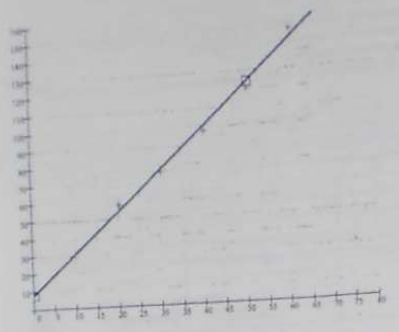
$$A=2{,}94,\quad B=7{,}27,\quad C=-1{,}25.$$

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 прямолинейной 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

$$r_в=\frac{\overline{XY}-\bar{X}\cdot\bar{Y}}{\sigma_x\,\sigma_y}=\frac{4496-40\cdot100{,}8}{14{,}14214\cdot33{,}9317}=0{,}96694.$$

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

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

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

Т.к. $t_{набл}>t_{крит}$ 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.

Номер испыт.	Количество (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
Средн.	85,8	78,6	68

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. последнюю строку таблицы).

Во-вторых, найдем коэффициенты R_j и P_j:

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

$R_1=91+80+95+88+75=429$; $P_1=91^2+80^2+95^2+88^2+75^2=37075$.
$R_2=74+50+90+85+94=393$; $P_2=74^2+50^2+90^2+85^2+94^2=32137$.
$R_3=70+61+55+76+78=340$; $P_3=70^2+61^2+55^2+76^2+78^2=23506$.

Найдём факторную дисперсию

$$s_{факт}^2=\frac{1}{q(l-1)}\left(\sum_{j=1}^{l}R_j^2-\frac{1}{l}\left(\sum_{j=1}^{l}R_j\right)^2\right)=\frac{1}{5\cdot2}\left((429^2+393^2+340^2)-\frac{1}{3}(429+393+340)^2\right)$$

Найдем остаточную дисперсию

$$s_{ост}^2=\frac{1}{l(q-1)}\left(\sum_{j=1}^{l}P_j-\frac{1}{q}\sum_{j=1}^{l}R_j^2\right)=\frac{1}{3\cdot4}\left((37075+32137+23506)-\frac{1}{5}(429^2+393^2+340^2)\right)=158.$$

Далее воспользуемся правилом:
а) если $s_{факт}^2<s_{ост}^2$ то следует сразу сделать вывод об отсутствии существенного влияния данного уровня фактора (количества вакцины) на процент заболевших людей.
б) в противном случае следует проверить значимость различия этих дисперсий при заданном уровне значимости и конкурирующей гипотезе $s_{факт}^2>s_{ост}^2$. Т.е. по таблице распределения Фишера-Снедекора (прил.6) определяем Критическое значение критерия

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

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 $s_{факт}^2=s_{ост}^2$, а потому следует сделать вывод о несущественности влияния данной вакцины на уровень заболеваемости (в случае $F_{набл}>F_{крит}$ делаем обратные выводы).

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

§ 2. Криволинейная корреляция

Если график регрессии изображается кривой линией, 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:

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

Аналогично находится выборочное уравнение регрессии X на Y

$$\bar{x}_y=A_1y^2+B_1y+C_1.$$

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

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

или (в других обозначениях)

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

Здесь

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

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

Y \ X	2	3	5	n_y
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

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

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