4. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

4.1. Парная корреляция

Основная задача корреляционного анализа – выявление значимости связи между значениями различных случайных величин. Величины, значения которых не зависят от того, какие значения получили некоторые другие величины, называются **независимыми** от них.

Зависимость между величинами (в том числе и случайными), при которых каждому значению одной величины (аргумента) отвечает одно или несколько вполне определенных значений другой, называют соответственно **однозначной** или **многозначной функциональной зависимостью**.

Зависимость между величинами, при которой каждому значению одной величины отвечает с соответствующей вероятностью множество возможных значений другой, называют **вероятностной** (стохастической, статистической). В общем случае вероятностной связи при изменении значения одной величины изменяется условный закон распределения другой.

Если при наличии вероятностной зависимости между двумя величинами с изменением значения одной величины изменяется только математическое ожидание второй (и наоборот), а дисперсия, области возможных значений и тип закона распределения остаются неизменными, то для таких величин характерна **корреляционная зависимость**.

Примерами корреляционной связи являются зависимости: между пределами прочности и текучести стали определенной марки, между погрешностью размера и погрешностью формы поверхности детали, обработанной определенным методом, между температурой ис-

пытания и ударной вязкостью стали, между усилием прижима ролика и шероховатостью накатанной детали. В первых двух примерах имеет место корреляционная связь между двумя откликами, а в третьем и четвертом – между фактором, который является случайной величиной в связи с погрешностью измерения, и откликом.

По характеру корреляционные связи могут быть прямолинейными и криволинейными.

Прямолинейной называется такая корреляционная связь, когда равным изменениям одной переменной соответствуют равные изменения другой переменной (рис. 4.1, *а, б*). В случае **криволинейной корреляции** равным изменениям одной переменной могут соответствовать любые изменения другой переменной (рис. 4.1, *в*).

Под **положительной корреляцией** подразумевается такая корреляция, когда с увеличением одной переменной увеличивается другая переменная (рис. 4.1, *а*). При **отрицательной корреляции** с увеличением одной переменной другая, наоборот, убывает (рис. 4.1, *б*).

На рис. 4.1, *г* представлен случай, когда между переменными отсутствует связь (нет корреляции).

Пусть X и Y – случайные переменные, имеющие нормальное распределение. Силу линейной статистической связи между ними можно оценить коэффициентом корреляции:

$$\rho = M\left[\frac{X - M(X)}{\sqrt{D(X)}}\right] \cdot \left[\frac{Y - M(Y)}{\sqrt{D(Y)}}\right]. \qquad (4.1)$$

Коэффициент корреляции принимает значения в интервале (– 1, +1) и не зависит от выбора начала отсчета и единиц величин X и Y. Чем больше отличается от нуля коэффициент корреляции, тем сильнее зависимость между величинами X и Y. Коэффициент корреляции независимых величин равен нулю. Однако обратное утверждение не все-

гда является верным, потому что на коэффициент корреляции оказывает влияние также отклонение от линейности связи.

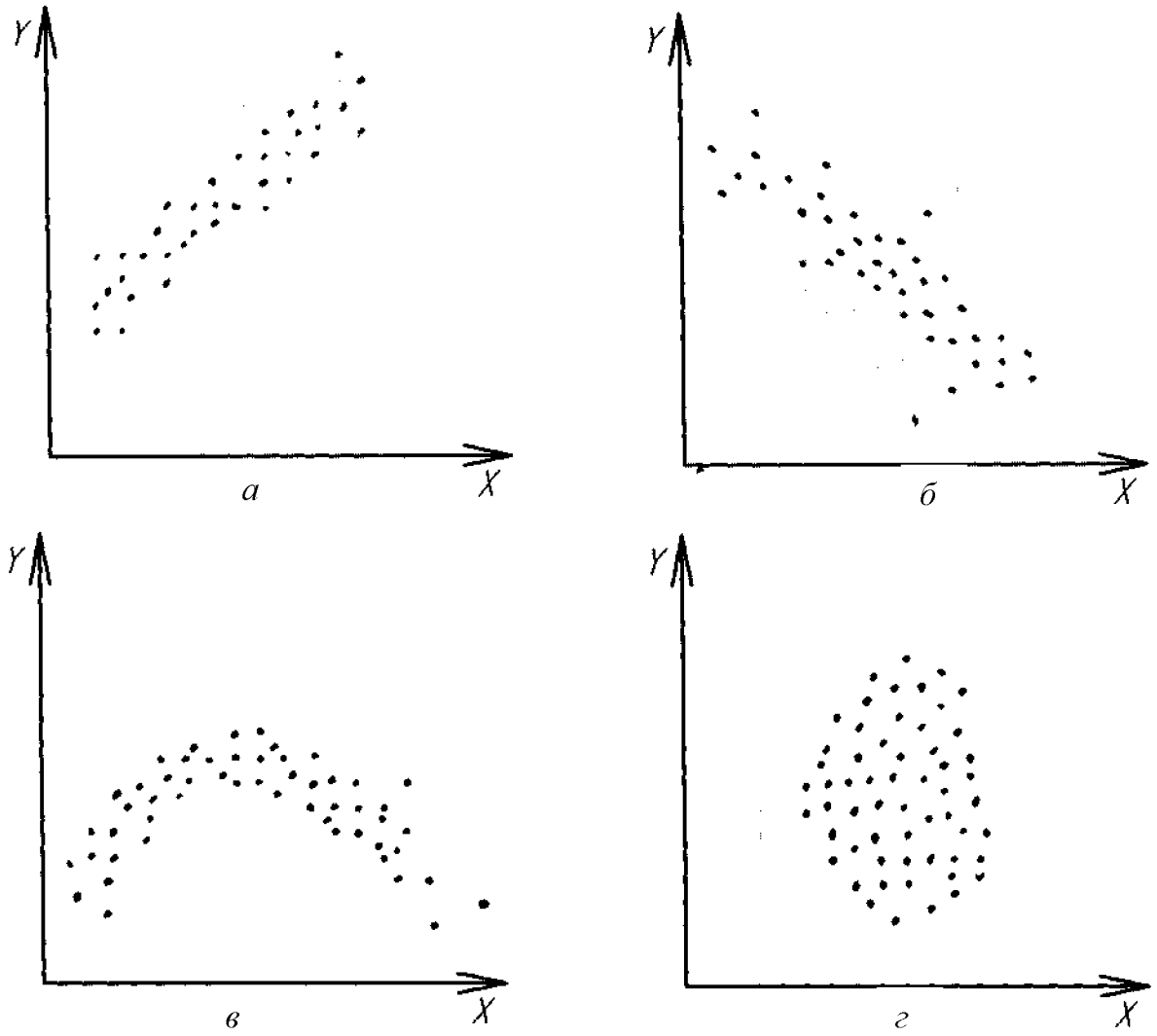

Рис. 4.1. Корреляционные зависимости

Характеристикой нелинейной корреляционной связи является корреляционное отношение

$$\eta_T^2 = \{M\,[M\,(Y/X = x) - M\,(Y)]^2\}/M\,[Y - M\,(Y)]^2, \qquad (4.2)$$

где $M\,(У / X = x)$ – условное математическое ожидание случайной переменной Y, рассматриваемое как функция x.

Оценкой коэффициента корреляции является значение коэффициента r. Для его вычисления необходимо знать оценки математических ожиданий $M(X)$ и $M(Y)$, а также дисперсий $D(X)$ и $D(Y)$. Если выполнено m наблюдений:

$$r = \left[\sum_{i=1}^{m}(X_i - \overline{X})(Y_1 - \overline{Y})\right] / \left[(m-1)S(X)S(Y)\right]. \tag{4.3}$$

При относительно небольшом значении m удобно пользоваться следующей системой формул:

$$\begin{cases} \sum_{i=1}^{m}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{m} X_i Y_i - \frac{1}{m}\sum_{i=1}^{m} X_i \sum_{i=1}^{m} Y_i; \\ (m-1)S^2(X) = \sum_{i=1}^{m} X_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} X_i\right)^2; \\ (m-1)S^2(Y) = \sum_{i=1}^{m} Y_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} Y_i\right). \end{cases} \tag{4.4}$$

На практике при небольших значениях m для вычисления выборочного коэффициента корреляции r используют поле корреляции и корреляционную таблицу.

Пару случайных чисел X_i, Y_i можно изобразить графически в виде точки с координатами (X_i, Y_i). По осям координат откладываются интервалы изменения переменных и наносится координатная сетка. Каждую пару переменных из данной выборки изображают точкой в соответствующей клетке. Такое изображение называют полем корреляции. На рис. 4.2 показано поле корреляции для 106 совместных измерений предела прочности σ_B и предела текучести σ_T стали 30ХГСА.

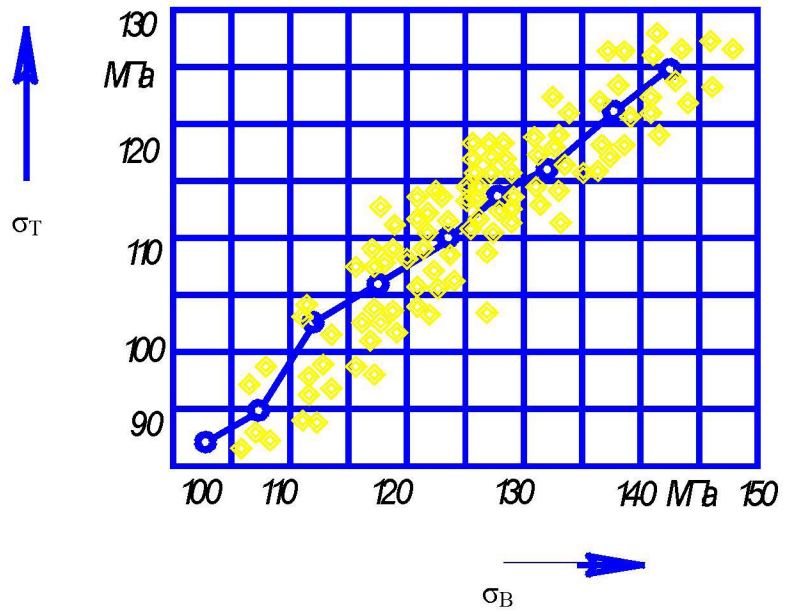

Рис. 4.2. Поле корреляции совместных измерений предела прочности σ_B и предела текучести σ_T стали 30ХГСА

Поле корреляции позволяет построить корреляционную таблицу (табл. 4.1).

В ячейки, образованные пересечением строк и столбцов, заносятся частоты m_{xy} попадания пар значений *(X, Y)* в соответствующие интервалы поля корреляции. В первом столбце и первой строке корреляционной таблицы указывают середины интервалов изменения случайных величин, а в последних – суммы частот m_{xy} по строкам и столбцам *(*m_x и m_y соответственно).

Система формул (4.4) может быть преобразована следующим образом:

$$\begin{cases} \sum_{i=1}^{m}(X_i-\overline{X})(Y_i-\overline{Y})=\sum_{i=1}^{m}x_i y_i m_{XE}-\frac{1}{m}\sum_{i=1}^{m}x_i m_X\sum_{i=1}^{m}y_i m_Y; \\ mS^2(X)=\sum_{i=1}^{m_1}x_i^2 m_X-\frac{1}{m}\left(\sum_{i=1}^{m_1}x_i m_X\right)^2; \\ mS^2(Y)=\sum_{i=1}^{m_2}y_i^2 m_Y-\frac{1}{m}\left(\sum_{i=1}^{m_2}y_i m_Y\right), \end{cases} \qquad (4.5)$$

где x_i и y_i – середины соответствующих интервалов изменения величин; m_1 – число столбцов; m_2 – число строк корреляционной таблицы; $m = \sum m_{XY} = \sum m_X = \sum m_Y$.

Таблица 4.1

Корреляционная матрица для σ_B и σ_T

σ_T / σ_B	925 (0)	975 (1)	1025 (2)	1075 (3)	1125 (4)	1175 (5)	1225 (6)	1275 (7)	Итого
1025 (2)	1								1
1075 (3)	1	1							2
1125 (4)		2	4	2					8
1175 (5)		1	7	11	1				20
1225 (6)			1	13	13				27
1275 (7)			1	2	15	9			27
1325 (8)					4	10	1		15
1375 (9)						1	3		4
1425 (10)							1	1	2
Итого	2	4	13	28	33	20	5	1	106

Если диапазоны изменения X и Y разделены на равные интервалы (Δx, Δy), то вместо натуральных значений можно использовать це-

лочисленные коды. Для этого нужно выполнить линейное преобразование x и y по формулам

$$x_{ki} = \frac{x_i - x_{\min}}{\Delta x}; \qquad y_{ki} = \frac{y_i}{\Delta y} - \frac{x_{\min}}{\Delta x}. \tag{4.6}$$

Тогда коды x_i примут значения x_{ki} = 0, 1, 2, ..., i, а коды y_{ki} будут сдвинуты на целое число $\Delta = y_{min}/\Delta y - x_{min}/\Delta x$.

Проверка значимости коэффициента корреляции производится при помощи критерия Стьюдента. Расчетное (наблюдаемое) значение критерия

$$t_{расч} = r\sqrt{(m-2)/(1-r^2)}. \tag{4.7}$$

Критическое (табличное) значение критерия определяется из таблицы приложения 2 по принятому значению доверительной вероятности и числу степеней свободы. Если $|t_{расч}| < t_{табл}$, то принимается нулевая гипотеза (p = 0). В противном случае она отклоняется.

Выборочное корреляционное отношение вычисляется по формуле

$$\eta^2 = S^2(Y/x) / S^2(Y), \tag{4.8}$$

где $S^2(Y)$ определяется по формуле (4.5); $S^2(Y/x)$ – условная дисперсия.

$$S^2(Y/x) = \frac{1}{m}\sum[\bar{y}(x) - \bar{y}]^2 m_X, \tag{4.9}$$

$$\bar{y}(x) = \frac{1}{m_X}\sum y_i m_{XY}. \tag{4.10}$$

Корреляционное отношение η_T^2 связано с ρ^2 следующим образом: $0 \le \rho^2 \le \eta_T^2 \le 1$. В случае линейной зависимости между переменными $\rho^2 = \eta_T^2$. Разность $\eta_T^2 - \rho^2$ может служить показателем нелинейной связи.

Пример. На основании данных корреляционной табл. 4.1 необходимо определить выборочный коэффициент корреляции r между σ_B и σ_T.

Кодом для σ_B выбираем x, а для σ_T – y. Тогда $\Delta = 2$. Коды записаны в корреляционной таблице под действительными значениями в скобках.

На основании формулы (4.5) получим:

$$\sum_{i=1}^{106}(X_i - \overline{X})(Y_i - \overline{Y}) = 2569 - \frac{1}{106}667 \cdot 384 = 152{,}698;$$

$$mS^2(X) = 4429 - \frac{1}{106}667^2 = 231{,}934;; \qquad S(x) = 1{,}479;$$

$$m\,S^2(Y) = 1566 - \frac{1}{106}384^2 = 174{,}906; \qquad S(y) = 1{,}285.$$

Согласно уравнению (4.3) $r = 152{,}698/(105 \cdot 1{,}479 \cdot 1{,}285) = 0{,}763$.

Пример. Для условий примера, представленного выше, определить корреляционное отношение. Значение $\overline{y}(x)$, вычисленное по формуле (4.10), для соответствующих значений кодов σ_B (первый столбец табл. 4.1):

$\overline{x}$	2	3	4	5	6	7	8	9	10
$\overline{y}(x)$	0	0,5	2	2,6	3,44	4,19	4,8	5,75	6,5

$$\bar{y} - \frac{1}{m}\sum y_i m_Y = 3{,}62; \quad S^2(Y \mid x) = 1{,}317; \quad \eta^2 = 1{,}317 / 1{,}285^2 = 0{,}798;$$

$r^2 = 0{,}763^2 = 0{,}582$. Характеристика нелинейности в данном случае $\eta^2 - r^2 = 0{,}798 - 0{,}582 = 0{,}216$.

Если соединить значения $\bar{y}(x)$, обозначенные на рис. 4.2 светлыми точками, штриховой линией, то будет видно, что полученная ломаная линия незначительно отклоняется от прямой.

4.2. Многомерный корреляционный анализ

Если имеется многомерная нормально распределенная совокупность с n признаками X_1, X_2, ..., X_n, то взаимозависимость между ними описывается корреляционной матрицей, под которой понимают матрицу, составленную из парных коэффициентов корреляции:

$$Q_n = \begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1л} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2k} & \dots & \rho_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \rho_{j1} & \rho_{j2} & \rho_{jk} & \dots & \rho_{jk} & \dots & \rho_{jk} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \rho_{n1} & \rho_{n2} & \rho_{nk} & \dots & \rho_{nk} & \dots & 1 \end{vmatrix},$$

где ρ_{jk} – парные коэффициенты корреляции.

Оценку парного коэффициента корреляции находят по формулам (4.3)–(4.5), в которые вместо x_i подставляют x_{ij}, а вместо $y_i - x_{ki}$.

В случае многомерной корреляции нельзя ограничиться одной корреляционной матрицей, так как зависимости между признаками сложны. Для более детального анализа используются частные коэффи-

циенты корреляций различных порядков, позволяющие оценивать связь между двумя признаками при фиксированном значении остальных.

Если исходная совокупность состоит из n признаков, то частный коэффициент корреляции l-го порядка отражает зависимость между двумя из них при фиксированных значениях l признаков из $(n-2)$ оставшихся. Если имеется, например, система из трех признаков X_1, X_2 и X_3, то можно определять частные коэффициенты корреляции только первого порядка, так как в данном случае нельзя фиксировать больше одного признака. Если фиксировать значение X_3, то

$$\rho_{12,3} = (\rho_{12} - \rho_{13}\rho_{23}) / \sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}, \qquad (4.11)$$

где $\rho_{12,3}$ – частный коэффициент корреляции между признаками X_1 и X_2 при фиксированном значении X_3; ρ_{12} , ρ_{13}, ρ_{23} – парные коэффициенты корреляции.

Аналогично определяются $\rho_{13,2}$ и $\rho_{23,1}$.

Расчет частных коэффициентов позволяет оценить взаимное влияние признаков. Если, например, корреляция между X_1 и X_2 основана только на общем влиянии X_3, то $\rho_{12,3} = 0$. Если имеются четыре признака, то можно зафиксировать значения одного или двух признаков. В последнем случае:

$$\rho_{12,34} = \frac{\rho_{12,4} - \rho_{13,4}\rho_{23,4}}{\sqrt{(1-\rho_{13,4}^2)(1-\rho_{23,4}^4)}} = \frac{\rho_{12,3} - \rho_{14,3}\rho_{24,3}}{\sqrt{(1-\rho_{14,3}^2)(1-\rho_{24,3}^2)}}. \qquad (4.12)$$

Частные коэффициенты корреляции вычисляются на основании оценок парных коэффициентов корреляции. Так же, как и для парных, проверяется значимость частных коэффициентов корреляции, но при этом число степеней свободы при исключении каждого признака

уменьшается на единицу. Если фиксируется значение одного признака, то в формулу (4.7) вместо (m – 2) следует подставить (m – 3), а если фиксируется значение двух признаков, – (m – 4).

Для оценки линейной связи одного из признаков со всеми остальными используется множественный, или совокупный, коэффициент корреляции. Для случая трех признаков коэффициент множественной корреляции оценивается по формуле

$$R_{123} = \sqrt{(r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23})/(1 - r_{23}^2)},$$

при этом X_2 и X_3, зависимость признака X_1 от которых оценивает R_{123}, считаются независимыми.

Для случая четырех признаков

$$R_{1234}^2 = 1 - (1 - r_{12}^2)(1 - r_{13,2}^2)(1 - r_{14,23}^2).$$

Значимость коэффициента множественной корреляции определяется при помощи критерия Фишера (F-критерия). Расчетное (наблюдаемое) значение критерия

$$F_{расч} = \frac{R^2}{1 - R^2} \cdot \frac{m - l - 2}{l},$$

где m – число наблюдений; $l = n - 1$; n – число признаков.

Согласно принятым доверительной вероятности и числу степеней свободы выбираем табличное значение коэффициента Фишера.

Пример. При изучении бесцентрового шлифования роликов методом на проход контролировались следующие показатели качества: погрешность формы X_1, погрешность размера X_3 и параметр шерохо-

ватости X_2. Всего было прошлифовано 50 образцов. При этом получена следующая матрица парных коэффициентов корреляции:

$$q_3 = \begin{pmatrix} 1 & 0{,}85 & 0{,}62 \\ & 1 & 0{,}53 \\ & & 1 \end{pmatrix}.$$

Необходимо установить характер взаимовлияния признаков. Согласно (4.11) имеем

$$r_{12,3} = (0{,}85 - 0{,}62 \cdot 0{,}53) / \sqrt{(1-0{,}62^2)(1-0{,}53^2)} = 0{,}784;$$

$$r_{23,1} = (0{,}53 - 0{,}85 \cdot 0{,}62) / \sqrt{(1-0{,}85^2)(1-0{,}62^2)} = 0{,}007;$$

$$r_{31,2} = (0{,}62 - 0{,}85 \cdot 0{,}53) / \sqrt{(1-0{,}85^2)(1-0{,}53^2)} = 0{,}379.$$

Проверяем значимости коэффициентов корреляции при помощи критерия Стьюдента (формула (4.7)). Расчетное (наблюдаемое) значение критерия:

$$t_{расч\ 1} = 0{,}784\sqrt{47/(1-0{,}784^2)} = 8{,}6584;$$

$$t_{расч\ 2} = 0{,}007\sqrt{47/(1-0{,}007^2)} = 0{,}048;$$

$$t_{расч\ 3} = 0{,}379\sqrt{47/(1-0{,}379^2)} = 2{,}808.$$

Согласно таблице приложения 2, при вероятности 0,95 , числе степеней свободы 60 $t_{табл} = 2{,}00$. Поскольку $t_{расч\ 1} > t_{табл}$ и $t_{расч\ 3} > t_{табл}$,

то нулевая гипотеза о незначимости связи между X_1 и X_2, а также X_1 и X_3 отвергается. Но $t_{расч\ 2} < t_{табл}$, поэтому величины X_2 и X_3 являются независимыми. Следовательно, условия шлифования, обусловливающие изменение погрешностей формы, влияют одновременно на уровень значения параметра шероховатости и погрешности размера. Причем эти условия сильнее влияют на шероховатость, чем на погрешность размера, так как $r_{12,3} > r_{31,2}$.

4.3. Корреляционные уравнения

Получение корреляционных уравнений – заключительный этап исследования связей между случайными величинами. Корреляционные уравнения позволяют вычислить вероятные значения одной случайной величины в зависимости от значений других случайных величин. Вероятным значением случайной величины Y называется ее значение, вычисленное с помощью корреляционного уравнения и близкое к условному математическому ожиданию $M(Y/X_i = X_i)$.

Корреляционное уравнение удобнее всего записывать в виде разложения по ортогональным многочленам Чебышева, что позволяет последовательно уточнять математическую модель с вычислением ошибки аппроксимации корреляционного уравнения полиномом данной степени.

Вначале корреляционная модель предполагается линейной:

$$(Y - \overline{Y})S(Y) = r\varepsilon, \qquad (4.13)$$

где $\varepsilon = (X - \overline{X}) / [\Delta X \cdot S(X)]$; r – оценка коэффициента корреляции между Y и X; $S(X)$, $S(Y)$ – оценки стандартного отклонения случайных величин X и Y; $\overline{X}$, $\overline{Y}$ – оценки математического ожидания величин; ΔX – шаг разбиения интервала изменения случайной величины X (см. формулу (4.6)).

Значения *S(X)* и *S(Y)* определяются из (4.5), *r* – из (4.3), при условии, что

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{k} x_i m_i \ ;$$

$$\overline{Y} = \frac{1}{m}\sum_{j=1}^{l} y_j m_j \ ,$$

где *k* и *l* – соответственно количество строк и столбцов корреляционной таблицы; m_i и m_j – суммы частот m_{XY} соответственно по *i*-й строке и *j*-му столбцу; *m* – общее число наблюдений; x_i и y_j – середины интервалов значений случайных величин.

Оценка ошибки определения *Y* с помощью корреляционного уравнения осуществляется по остаточной дисперсии:

$$S_0^2 = S^2(Y) \cdot (1 - r^2).$$

Проверка линейности связи между Y и X производится при помощи критерия линейности:

$$K_1 = \eta^2 - r^2,$$

который вычисляется с ошибкой:

$$S(K_1) = \sqrt{K_1 / m} \ .$$

Значение η^2 рассчитывается по формулам (4.8) и (4.9). Если K_1 незначимо отличается от $S(K_1)$, то можно остановиться на линейной модели ((формула (4.13)). Если K_1 значимо отличается от нуля, то

корреляционное уравнение предполагается в виде полинома второй степени:

$$(Y - \overline{Y}) / S(Y) = r\varepsilon + \frac{b_1}{a_1}(\varepsilon^2 + r_{30}\varepsilon - 1),$$

где $a_1 = r_{40} - r_{30}^2 - 1$; $b_1 = r_{21} - r_{30}r$;

$$r_{30} = \frac{\mu_{30}}{S^3(X)}; \quad r_{40} = \frac{\mu_{40}}{S^4(X)}; \quad r_{21} = \frac{\mu_{21}}{S^2(X)S(Y)};$$

$$\mu_{30} = M_{30} - 3\mu_{20}M_{10} - M_{10}^3; \quad \mu_{20} = M_{20} - M_{10}^2 = S^2(X);$$

$$\mu_{40} = M_{40} - 2\mu_{11}M_{10} - 6\mu_{20}M_{10}^2 - M_{10}^4; \quad \mu_{21} = M_{21} - 2\mu_{11}M_{10} - M_{20}M_{01};$$

$$\mu_{11} = M_{11} - M_{10}M_{01}; \quad M_{10} = \overline{X}; \quad M_{01} = \overline{Y};$$

$$M_{20} = \frac{1}{m}\sum_{i=1}^{k} x_i^2 m_i; \quad M_{30} = \frac{1}{m}\sum_{i=1}^{k} x_i^3 m_i;$$

$$M_{40} = \frac{1}{m}\sum_{i=1}^{k} x_i^4 m_i; \quad M_{02} = \frac{1}{m}\sum_{j=1}^{l} x_j^2 m_j;$$

$$\mathrm{M}_{11} = \frac{1}{m}\sum_{j=1}^{l}\sum_{i=1}^{k} x_i y_j m_{XY}; \quad \mathrm{M}_{21} = \frac{1}{m}\sum_{j=1}^{l}\sum_{i=1}^{k} x_i^2 y_i m_{XY}.$$

Погрешность определения Y с помощью корреляционного уравнения второго порядка оценивается остаточной дисперсией:

$$S_0^2 = S^2(Y)\cdot(1 - r^2 - b_1^2 / a_1).$$

Проверка квадратичности связи между X и Y производится при помощи критерия квадратичности:

$$K_2 = K_1 - b_1^2 / a_1,$$

который вычисляется с ошибкой:

$$S(K_2) = \sqrt{K_2 / m}\,.$$

Если K_2 значимо отличается от нуля, то корреляционное уравнение можно представить в виде полинома второй степени, и т. п.