§4.1. Линейная корреляционная зависимость

Часто на практике требуется установить вид и оценить силу зависимости изучаемой случайной величины Y от одной или нескольких других величин (случайных или неслучайных). Рассмотрим сначала зависимость случайной величины Y от одной случайной X. Две величины могут быть связаны:
1) функциональной зависимостью (Y=f(X)), либо 2) статистической зависимостью.

Статистическая зависимость – зависимость, при которой изменение одной из величин влечет изменение распределения другой (вида распределения, либо числовых характеристик распределения).

Корреляционная зависимость – статистическая зависимость, при которой изменение одной из величин влечет изменение среднего значения другой. С математической точки зрения корреляционная зависимость – функциональная зависимость условного среднего $\bar{y}_x$ от x :

$$\bar{y}_x = f(x), \qquad (1)$$

где $\bar{y}_x$- выборочное условное среднее (среднее арифметическое значений Y, соответствующих значению x величины X); уравнение (1) называют выборочным уравнением регрессии Y на X; $f(x)$- выборочная функция регрессии Y на X; график функции $f(x)$ называют линией регрессии Y на X. Аналогично, $\bar{x}_y$ - условное среднее X на Y; $\bar{x}_y = \varphi(y)$- выборочное уравнение регрессии X на Y; $\varphi(y)$- функция регрессии X на Y; график функции $\varphi(y)$ называют линией регрессии X на Y.

Задачи теории корреляции

Теория корреляции решает следующие задачи:

1) Установление формы корреляционной зависимости, т.е. вида функций $f(x), \varphi(y)$(если обе функции $f(x), \varphi(y)$ являются линейными, то корреляционная зависимость называется линейной; в противном случае – нелинейной корреляционной зависимостью);
2) Оценка силы (тесноты) корреляционной зависимости.

Пусть в результате независимых испытаний получено n пар значений (x_i, y_i). Предположим, что X и Y связаны линейной корреляционной зависимостью, т.е. $f(x) = \alpha x + \beta$.

Найдем по выборочным значениям (x_i, y_i) точечные оценки параметров α, β так, чтобы точки (x_i, y_i), построенные на координатной плоскости, находились вблизи прямой

$$\bar{y}_x = \alpha x + \beta. \qquad (1)$$

Метод наименьших квадратов

Выборочные параметры α, β находят из условия обращения в минимум функции

$$Q(b_{yx}, \bar{b}) = \sum_{i=1}^{n} (f(x_i) - y_i)^2 = \sum_{i=1}^{n} (b_{yx} x_i + \bar{b} - y_i)^2$$

Для отыскания минимума функции

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (f(x_i) - y_i)^2 = \sum_{i=1}^{n} (\alpha x_i + \beta - y_i)^2.$$

приравняем нули соответствующие частные производные

$$\frac{\partial Q}{\partial \alpha} = 2\sum_{i=1}^{n} (\alpha x_i + \beta - y_i) x_i = 0, \qquad (2)$$

$$\frac{\partial Q}{\partial \beta} = 2\sum_{i=1}^{n} (\alpha x_i + \beta - y_i) = 0.$$

Выполняя элементарные преобразования, получим систему двух линейных уравнений относительно α, β

$$\begin{cases} \overline{x^2}\alpha + \bar{x}\beta = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} x_i y_j, \\ \bar{x}\alpha + \beta = \bar{y}, \end{cases} \qquad (3)$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, n_{ij} - число наблюдений одной и той же пары значений (x_i, y_i), k- число различных пар (x_i, y_i). Из системы уравнений (3) следует, что $\beta = \bar{y} - \alpha\bar{x}$,

$$\alpha = \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} x_i y_j - n\bar{x}\bar{y}}{nS_x^2}, \qquad S_x^2 = \overline{x^2} - \bar{x}^2.$$

При этом выборочное уравнение регрессии Y на X примет вид

$$\bar{y}_x - \bar{y} = r\frac{S_y}{S_x}(x - \bar{x}),$$

где $r = \alpha\frac{S_x}{S_y}$ - выборочный коэффициент корреляции. Аналогично уравнение регрессии X на Y имеет вид

$$\bar{x}_y - \bar{x} = r\frac{S_x}{S_y}(y - \bar{y}), \text{ где } r = \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} x_i y_j - n\bar{x}\bar{y}}{nS_x S_y}.$$

Пример. При большом числе наблюдений одно и то же значение x может встретиться n_x раз, одно и то же значение y может встретиться n_y раз, одна и та же пара значений чисел (x,y) может

наблюдаться n_{xy} раз. Поэтому данные наблюдений группируют, т.е. подсчитывают n_x, n_y, n_{xy}. Все сгруппированные данные записывают в виде таблицы, которую называют корреляционной.

X / Y	10	20	30	40	50	60	n_y
15	5	7					12
25		20	23				43
35			30	47	2		79
45			10	11	20	6	47
55				9	7	3	19
n_x	5	27	63	67	29	9	$n = 200$

В первой строке таблицы указаны наблюдаемые значения величины X, а в первом столбце – наблюдаемые значения величины Y. На пересечении строк и столбцов вписаны частоты n_{xy} наблюдаемых пар значений этих величин. Например, частота 5 указывает, что пара чисел (10,15) наблюдалась 5 раз. В последнем столбце записаны суммы частот строк. В последней строке записаны суммы частот столбцов.

Вычислим выборочный коэффициент корреляции по данным корреляционной таблицы. Можно значительно упростить вычисления, если перейти к условным вариантам $u_i = \frac{x_i - c_1}{h_1}$, $v_i = \frac{y_i - c_2}{h_2}$, переход к которым не меняет величины выборочного коэффициента корреляции

$$r = \frac{\sum n_{xy} xy - n\bar{x}\,\bar{y}}{nS_x S_y} = \frac{\sum n_{uv} uv - n\bar{u}\,\bar{v}}{nS_u S_v}$$

В данном примере $u_i = \frac{x_i - c_1}{h_1} = \frac{x_i - 40}{10}$, где в качестве нуля c_1 взята варианта $x = 40$, имеющая наибольшую частоту 47; шаг h_1 равен разности между двумя соседними вариантами. Условные варианты $v_i = \frac{y_i - c_2}{h_2} = \frac{y_i - 35}{10}$, где в качестве нуля c_2 взята варианта $y = 35$, имеющая наибольшую частоту 47; шаг h_2 равен разности между двумя

u v	-3	-2	-1	0	1	2	n_v
-2	5	7					12
-1		20	23				43
0			30	47	2		79
1			10	11	20	6	47
2				9	7	3	19
n_u	5	27	63	67	29	9	$n = 200$

соседними вариантами. Составим корреляционную таблицу в условных вариантах. Практически это делается так: в первом столбце вместо варианты 35, имеющей наибольшую частоту, пишут 0; над нулем пишут последовательно –1,-2…; под нулем пишут 1,2…. В первой строке вместо варианты 40, имеющей наибольшую частоту, пишут 0; слева от нуля последовательно пишут –1,-2….; справа от нуля пишут 1,2…. Все остальные данные переписывают из первоначальной корреляционной таблицы. В итоге получим корреляционную таблицу в условных вариантах.

Найдем $\overline{u}$ и $\overline{v}$

$$\overline{u} = \frac{\sum n_u u}{n} = \frac{5\cdot(-3) + 27\cdot(-2) + 63\cdot(-1) + 29\cdot 1 + 9\cdot 2}{200} = -0.425,$$

$$\overline{v} = \frac{\sum n_v v}{n} = \frac{12\cdot(-2) + 43\cdot(-1) + 47 + 19\cdot 2}{200} = 0.090$$

Вычислим вспомогательную величину $\overline{u^2}$, а затем S_u:

$$\overline{u^2} = \frac{\sum n_u u^2}{n} = \frac{5\cdot 9 + 27\cdot 4 + 1\cdot 63 + 1\cdot 29 + 9\cdot 4}{200} = 1.405$$

$$S_u = \sqrt{\overline{u^2} - \overline{u}^2} = \sqrt{1.405 - 0.425^2} = 1.106$$

Аналогично получим $S_v = 1.209$.

Найдем $\sum n_{uv} uv$ метод 4 полей, для чего составим расчетную таблицу

U V	-3	-2	-1	0	1	2
-2	5	7				
-1		20	23			
0						
1					20	6
2					7	3

I	30	68	23	II		
III			-10	IV	34	24

Название метода связано с тем, что строка и столбец, пересекающиеся в клетке, содержащей наибольшую частоту, делят корреляционную таблицу на 4 части, которые называют полями. Поле нумеруется так, как указано в таблице.

Найдем произведения пар вариант u и v и поместим их в верхние правые углы клеток, содержащих соответствующие частоты. Заполнив подобным образом остальные клетки 1,2,3,4 полей, получим таблицу, приведенную выше. Сложив числа итоговых клеток, получим $\sum n_{uv}uv = 121 - 10 + 58 = 169$. Найдем искомый коэффициент корреляции

$$r = \frac{\sum n_{uv}uv - n\,\bar{u}\,\bar{v}}{n\,S_u S_v} = \frac{169 - 200(-0.425)\cdot 0.09}{200\cdot 1.106\cdot 1.209} = 0.603$$

Теперь, когда известно как вычисляют r уместно привести пример на отыскание уравнения прямой линии регрессии. Поскольку при нахождении r уже вычислены $\bar{u}, \bar{v}, S_u, S_v$, то для нахождения $\bar{x}, \bar{y}, S_x, S_y$ целесообразно вывести формулы, связывающие $\bar{u}, \bar{v}, S_u, S_v$ и $\bar{x}, \bar{y}, S_x, S_y$. Выведем эти формулы

$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - c_1}{h_1}\right) = \frac{\bar{x}}{h_1} - \frac{c_1}{h_1}$, так что $\bar{x} = \bar{u}h_1 + c_1$. Аналогично $\bar{y} = \bar{v}h_2 + c_2$.

Тогда $S_u = \sqrt{\frac{1`}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - c_1}{h_1} - \frac{\bar{x} - c_1}{h_1}\right)^2} = \frac{1}{h_1}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_x}{h_1}$,

так что $S_x = h_1 S_u$. Аналогично $S_y = h_2 S_v$. Итак

$$\bar{x} = \bar{u}h_1 + c_1,\ \bar{y} = \bar{v}h_2 + c_2,\ S_x = h_1 S_u,\ S_y = h_2 S_x$$

Запишем искомое уравнение в общем виде

$$\bar{y}_x - \bar{y} = r\frac{S_y}{S_x}(x - \bar{x}) \qquad (4)$$

Коэффициент корреляции уже ранее вычислен и равен r = 0.603. Остается найти $\bar{x}, \bar{y}, S_x, S_y$:

$$\begin{aligned} \bar{x} &= \bar{u}h_1 + c_1 = -0.425 \cdot 10 + 40 = 35.75 \\ \bar{y} &= \bar{v}h_2 + c_2 = 0.09 \cdot 10 + 35 = 35.9 \\ S_x &= S_u h_1 = 1.106 \cdot 10 = 11.06 \\ S_y &= S_v h_2 = 1.209 \cdot 10 = 12.09 \end{aligned} \qquad (5)$$

Подставляя (5) в (4), получим искомое уравнение

$$\bar{y}_x - 35.9 = 0.603 \cdot \frac{12.09}{11.06}(x - 35.75)$$

или окончательно

$$\bar{y}_x = 0.659 \cdot x + 12.34 \qquad (6)$$

Сравним условные средние, вычисленные по уравнению (6) и по данным корреляционной таблицы. Например, при x=30: по уравнению (6) получим

$$\bar{y}_{30} = 0.659 \cdot 30 + 12.34 = 32.11,$$

а по таблице $\bar{y}_{30} = \dfrac{23 \cdot 25 + 30 \cdot 35 + 10 \cdot 45}{63} = 32.94.$

Как видим, согласование расчетного (согласно (6)) и наблюдаемого условных средних – удовлетворительное.

Доверительные оценки параметров прямой регрессии y на x.

При нахождении доверительного интервала для оценки параметров α, β теоретической прямой линии регрессии y на x используется сумма квадратов отклонений измеренных значений y_i от рассчитанных по выборочному уравнению прямой линии регрессии:

$$Q = \sum_{i=1}^{n} \left[y_i - r\frac{S_y}{S_x}(x_i - \bar{x}) \right]^2 = (n-1)\left(1 - r^2\right)S_y^2 .$$

При этом предполагается, что все ошибки измерения независимы и одинаково распределены по нормальному закону с центром 0 и дисперсией σ^2.

Границы доверительного интервала для параметра β равны

$$\bar{y} \pm t\sqrt{\frac{n-1}{n-2}}\sqrt{1-r^2}\,\frac{S_y}{\sqrt{n}}$$

а границами доверительного интервала для параметра α служат

$$\alpha \pm t\frac{S_y}{S_x}\sqrt{\frac{1-r^2}{n-2}},$$

где коэффициент t берется из таблицы распределения Стьюдента при числе степеней свободы $k = n-2$.

Доверительный интервал для оценки отклонения теоретической прямой линии регрессии от эмпирической

При фиксированном значении $x = x_0$ границы доверительного интервала для теоретической прямой регрессии определяются формулами

$$y_t(x_0) = y_e(x_0) \pm \frac{t}{\sqrt{n-2}}\sqrt{1+\frac{n(x_0-\bar{x})^2}{(n-1)S_x^2}},$$

здесь $y_e(x_0) = \bar{y} + \alpha(x_0 - \bar{x})$, коэффициент t берется из таблицы распределения Стьюдента при числе степеней свободы $k = n-2$. Следует помнить, что эта оценка значительно ухудшается по мере удаления от среднего значения $\bar{x}$.

Например, для вышеприведенного примера $t(0.95,198) = 1.96$ и соответственно границы доверительного интервала для $x_0 = 30$ равны 32.11 ± 1.30, так что наблюдаемое среднее $\bar{y}_{30} = 32.94$ принадлежит доверительному интервалу.

Свойства выборочного коэффициента корреляции

Выведем формулы

$$S_{\bar{y}x}^2 = S_y^2 \cdot (1-r^2) \qquad (1)$$

$$S_{\bar{x}y}^2 = S_x^2 \cdot (1-r^2) \qquad (2)$$

Для этого предположим, что величины Y и X связаны линейной корреляционной зависимостью

$$\bar{y}_x = \alpha + \beta x$$

Тогда получим

$$S^2_{\bar{y}_x} = \frac{\Sigma(y_i - \bar{y}_x)^2}{n} = \frac{\Sigma[y_i - \alpha - \beta x_i]^2}{n} = \frac{\Sigma[(y_i - \bar{y}) - \beta(x_i - \bar{x}) + (\bar{y} - \beta\bar{x} - \alpha)]^2}{n} =$$

$$= \frac{\Sigma(y_i - \bar{y})^2}{n} + \beta^2 \frac{\Sigma(x_i - \bar{x})^2}{n} + (\bar{y} - \beta\bar{x} - \alpha)^2 - 2\beta \frac{\Sigma(y_i - \bar{y}) \cdot (x_i - \bar{x})}{n} +$$

$$+ 2(\bar{y} - \beta\bar{x} - \alpha)\frac{\Sigma(y_i - \bar{y})}{n} + 2(\bar{y} - \beta\bar{x} - \alpha)\frac{\Sigma(x_i - \bar{x})}{n} = S_y^2 + \beta^2 S_x^2 + (\bar{y} - \beta\bar{x} - \alpha)^2 -$$

$$- 2\beta \frac{\Sigma xy - n \cdot \bar{x} \cdot \bar{y}}{n}$$

Итак, окончательно имеем

$$S^2_{\bar{y}_x} = S_y^2 + \beta^2 S_x^2 + (\bar{y} - \beta\bar{x} - \alpha)^2 - 2\beta \frac{\Sigma xy - n\bar{x} \cdot \bar{y}}{n} \qquad (3)$$

Параметры α, β найдем из условия минимума функции $S^2_{\bar{y}_x}(\alpha, \beta)$. Необходимые условия минимума этой функции имеют вид

$$\frac{\partial S^2_{\bar{y}_x}}{\partial \alpha} = -2(\bar{y} - \beta\bar{x} - \alpha) = 0 \qquad (4)$$

$$\frac{\partial S^2_{\bar{y}_x}}{\partial \beta} = 2\beta S_x^2 - 2\bar{x}(\bar{y} - \beta\bar{x} - \alpha) - 2\frac{\Sigma xy - n\bar{x}\,\bar{y}}{n} = 0 \qquad (5)$$

Из уравнения (5) находим

$$\beta = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{nS_x^2} = r\frac{S_y}{S_x} \qquad (6)$$

Из уравнения (4) следует

$$\alpha = -\bar{y} + \beta\bar{x} = -\bar{y} + r\frac{S_y}{S_x}\bar{x} \qquad (7)$$

Подставляя (6) и (7) в (3), получим формулу (1) $S^2_{\bar{y}_x} = S_y^2 \cdot (1 - r^2)$. Аналогично $S^2_{\bar{x}_y} = S_x^2 \cdot (1 - r^2)$.

Свойство 1. Выборочный коэффициент корреляции по модулю не превосходит единицу $-1 \le r \le 1$.

Доказательство

Так как любая дисперсия неотрицательна, т.е. $S^2_{\bar{y}_x} \ge 0, S_y^2 \ge 0$, то из формулы $S^2_{\bar{y}_x} = S_y^2 \cdot (1 - r^2)$ следует, что $r^2 \le 1$ или $-1 \le r \le 1$, что и требовалось показать.

Свойство 2. Если $r = 0$, то наблюдаемые значения x, y не связаны линейной корреляционной зависимостью.

Доказательство

Доказательство проведем по методу от противного. Предположим, что наблюдаемые значения x, y связаны линейной корреляционной зависимостью, т.е.

$$\bar{y}_x = \bar{y} + b_{yx}(x - \bar{x})$$
$$\bar{x}_y = \bar{x} + b_{xy}(y - \bar{y})$$

Отсюда при $r = 0$ следует $\bar{y}_x = \bar{y}$, $\bar{x}_y = \bar{x}$, что противоречит предположению.

Замечание. Если $r = 0$, то x, y могут быть связаны нелинейной корреляционной зависимостью или даже функциональной зависимостью.

Свойство 3. Если $|r| = 1$, то наблюдаемые значения x, y связаны линейной функциональной зависимостью.

Доказательство

При $|r| = 1$ из формулы $S^2_{\bar{y}_x} = S^2_y \cdot \left(1 - r^2\right)$ следует, что $S^2_{\bar{y}_x} = \dfrac{\Sigma\left(y_i - \bar{y}_x\right)^2}{n} = 0$, т.е. $y_i = \bar{y}_x$. Тогда из $\bar{y}_x = \bar{y} + b_{yx}(x - \bar{x})$ следует $y_i = \bar{y} + b_{yx}(x_i - \bar{x})$, что и требовалось доказать.

Замечание. Из свойства (3) следует, что только наблюдаемые значения, а не все возможные значения связаны линейной функциональной зависимостью.

Из доказанных свойств следует, что r характеризует силу линейной корреляционной зависимости между количественными признаками в выборке:

1) чем ближе $|r|$ к единице, тем связь сильнее;
2) чем ближе $|r|$ к нулю, тем связь слабее.

Замечание. Если выборка имеет достаточно большой объем, то заключение о силе линейной корреляционной зависимости между наблюдаемыми значениями признаков может быть распространена на всю совокупность значений признаков X и Y.

§4.2. Выборочное корреляционное отношение

Для оценки тесноты линейной корреляционной связи между физическими величинами в выборке служит выборочный коэффициент корреляции r. Для оценки тесноты любой корреляционной связи вводят другие характеристики.

Пусть данные наблюдений за количественными признаками X и Y сведены в корреляционную таблицу. Тем самым наблюдаемые

значения Y оказываются разбиты на группы; каждая группа содержит те значения Y, которые соответствуют определенному значению X. Так как все значения признака Y разбиты на группы, то можно представить

$$D_{общ} = D_{внгр} + D_{межгр} \qquad (1)$$

При этом оказывается справедливым следующее утверждение.

Утверждение12. 1)Если величина Y связана с величиной X функциональной зависимостью, то

$$\frac{D_{межгр}}{D_{общ}} = 1$$

2)если величина Y связана с величиной X корреляционной зависимостью, то

$$\frac{D_{межгр}}{D_{общ}} < 1$$

Докажем это утверждение. Доказательство разобьем на две части. Сначала докажем первую часть утверждения.

1) Если случайная величина Y связана с случайной величиной X функционально, то по определению функциональной зависимости определенному значению x соответствует только одно значение y. Поэтому в каждой j группе ее элементы равны между собой, т.е.

$$y_{1j} = y_{2j} = ... = y_j^* \qquad (2)$$

Из (2) следует, что групповое среднее

$$\bar{y}_j = \frac{1}{N_j}\sum_{i=1}^{m_j} n_{ij} y_{ij} = \frac{1}{N_j}\sum_{i=1}^{m_j} n_{ij} y_j^* = \frac{y_j^*}{N_j}\sum_{i=1}^{m_j} n_{ij} = y_j^* \qquad (3)$$

Следовательно, групповая дисперсия равна

$$D_j = \frac{1}{N_j}\sum_{i=1}^{m_j} n_{ij}\left(y_{ij} - \bar{y}_j\right)^2 = \frac{1}{N_j}\sum_{i=1}^{m_j} n_{ij}\left(y_j^* - y_j^*\right)^2 = 0 \qquad (4)$$

В свою очередь, из (4) вытекает, что

$$D_{внгр} = \frac{1}{n}\sum_{j=1}^{k} N_j D_j = 0 \qquad (5)$$

Подставляя (5) в (1), получим

$$D_{общ} = D_{межгр}$$

Отсюда находим $\frac{D_{межгр}}{D_{общ}} = 1$, что и требовалось показать в первой части.

2) Если случайная величина Y связана с случайной величиной X корреляционной зависимостью, то определенному значению x

соответствуют, вообще говоря, различные значения y_{ij}, образующие группу. Поэтому в этом случае $D_j \neq 0$. Следовательно, $D_{внгр} \neq 0$. Так как $D_{внгр} > 0,\ D_{межгр} \geq 0$, то $D_{межгр} < D_{межгр} + D_{внгр}$, т.е. согласно (1) $D_{межгр} < D_{общ}$, так что $\frac{D_{межгр}}{D_{общ}} < 1$, что и требовалось показать во второй части.

Из доказанного утверждения видно, что чем связь между величинами ближе к функциональной, тем больше будет приближаться к единице отношение $\frac{D_{межгр}}{D_{общ}}$. Отсюда ясно, что целесообразно рассматривать в качестве меры тесноты корреляционной зависимости отношение $\frac{D_{межгр}}{D_{общ}}$ или, что то же, отношение $\frac{S_{межгр}}{S_{общ}}$.

Выборочным корреляционным отношением Y к X называют отношение вида

$$\eta_{yx} = \frac{S_{\bar{y}_x}}{S_y},$$

где $S_{\bar{y}_x} = \sqrt{D_{межгр}} = \sqrt{\frac{1}{n}\sum n_x(\bar{y}_x - \bar{y})^2}$, $S_y = \sqrt{D_{общ}} = \sqrt{\frac{1}{n}\sum n_y(y - \bar{y})^2}$,

n – объем выборки, n_x - частота значения x случайной величины X; n_y - частота значения y случайной величины Y; $\bar{y}$ - общее среднее величины Y; $\bar{y}_x$ - групповое среднее величины Y. Аналогично определяется выборочное корреляционное отношение X к Y

$$\eta_{xy} = \frac{S_{\bar{x}_y}}{S_x}.$$

Пример. По данным корреляционной таблицы вычислим η_{yx}.

X / Y	10	20	30	n_y	$\bar{x}_y$
15	4	28	6	38	20.5
25	6		6	12	20
n_x	10	28	12	n=50	
$\bar{y}_x$	21	15	20		

Найдем общее среднее

$$\bar{y} = \frac{1}{n}\sum n_y y = \frac{38\cdot 15 + 12\cdot 25}{50} = 17.4$$

Вычислим выборочные средние квадратичные отклонения

$$S_y = \sqrt{\frac{1}{n}\sum n_y (y - \bar{y})^2} = \sqrt{\frac{38(15 - 17.4)^2 + 12(25 - 17.4)^2}{50}} = 4.27.$$

$$S_{\bar{y}_x} = \sqrt{\frac{1}{n}\sum n_x (\bar{y}_x - \bar{y})^2} = \sqrt{\frac{10(21 - 17.4)^2 + 28(15 - 17.4)^2 + 12(20 - 17.4)^2}{50}} = 2.73$$

Тогда выборочное корреляционное отношение равно

$$\eta_{yx} = \frac{S_{\bar{y}_x}}{S_y} = \frac{2.73}{4.27} = 0.64.$$

Аналогично

$$\bar{x} = \frac{1}{n}\sum n_x x = \frac{10\cdot 10 + 28\cdot 20 + 12\cdot 30}{50} = 20.4,\ \bar{x}_{15} = \frac{4\cdot 10 + 28\cdot 20 + 6\cdot 30}{38} = 20.53$$

$$\bar{x}_{25} = \frac{6\cdot 10 + 6\cdot 30}{12} = 20$$

$$S_x = \sqrt{\frac{1}{n}\sum n_x (x - \bar{x})^2} = \sqrt{\frac{10(10 - 20.4)^2 + 12(30 - 20.4)^2 + 28(20 - 20.4)^2}{50}} = \frac{\sqrt{1096}}{5}.$$

$$S_{\bar{x}_y} = \sqrt{\frac{1}{n}\sum n_y (\bar{x}_y - \bar{x})^2} = \sqrt{\frac{38(20.5 - 20.4)^2 + 12(20 - 20.4)^2}{50}} = \frac{\sqrt{1.15}}{5}.$$

$$\eta_{xy} = \frac{S_{\bar{x}_y}}{S_x} = \sqrt{\frac{1.15}{1096}} = 0.33.$$

Видим, что, вообще говоря, $\eta_{yx} \neq \eta_{xy}$.

Вычислим теперь выборочный коэффициент корреляции r и сравним его с корреляционным отношением η. Для этого перейдем к условным вариантам

$$u_i = \frac{x_i - 20}{10},\ v_i = \frac{y_i - 15}{10}$$

Перепишем в условных координатах корреляционную таблицу

U / V	-1	0	1	n_v
0	4	28	6	38
1	6		6	12
n_u	10	28	12	n=50

Вычислим значение выборочного коэффициента корреляции

$$r = \frac{\sum n_{uv} uv - n\bar{u}\,\bar{v}}{n\sigma_u\sigma_v}$$

Для этого найдем соответствующие средние значения

$$\bar{u}=\frac{-1\cdot 10+0\cdot 28+1\cdot 12}{50}=\frac{2}{50},\ \bar{v}=\frac{0\cdot 38+1\cdot 12}{50}=\frac{12}{50},\ \overline{u^2}=\frac{22}{50},\ \overline{v^2}=\frac{12}{50}$$

Найдем выборочные средние отклонения

$$\sigma_u=\sqrt{\overline{u^2}-\bar{u}^2}=\frac{\sqrt{1096}}{50},\ \sigma_v=\sqrt{\overline{v^2}-\bar{v}^2}=\frac{\sqrt{1056}}{50}$$

Вычислим сумму $\sum n_{uv}uv=4\cdot 0+6\cdot(-1)+1\cdot 0\cdot 6+1\cdot 1\cdot 6=0$. Тогда значение выборочного коэффициента корреляции равно

$$r=-\frac{\bar{u}\,\bar{v}}{\sigma_u\,\sigma_v}=-\frac{2\cdot 12}{50\cdot 50\cdot\frac{\sqrt{1096}}{50}\cdot\frac{\sqrt{1056}}{50}}=-0.02$$

В данном примере $\eta>|r|$. Это соотношение является общим.

Свойства выборочного корреляционного отношения

Так как η_{yx} обладает теми же свойствами, что и η_{xy}, то рассмотрим свойства только η_{yx}, которое для упрощения записи обозначим через η и будем называть корреляционным отношением.

Свойство 1. $0\le\eta\le 1$

Доказательство

Так как по определению $D_{межгр}\ge 0,\ D_{общ}\ge 0$, то

$$S_{\bar{y}_x}=\sqrt{D_{межгр}}\ge 0,\ S_y=\sqrt{D_{общ}}\ge 0.$$

Следовательно, $\eta\ge 0$. Из соотношения

$$D_{общ}=D_{межгр}+D_{внгр}$$

следует, что

$$\frac{D_{внгр}}{D_{общ}}+\frac{D_{межгр}}{D_{общ}}=1 \quad \text{или} \quad \frac{D_{внгр}}{D_{общ}}+\eta^2=1$$

Так как $\frac{D_{внгр}}{D_{общ}}\ge 0,\ \eta^2\ge 0$, то каждое из слагаемых ≤ 1; в частности, $\eta^2\le 1$. Приняв во внимание, что $\eta\ge 0$, заключаем $0\le\eta\le 1$, что и требовалось показать.

Свойство 2. Если $\eta=0$, то Y и X корреляционной зависимостью не связаны.

Доказательство

Из $\eta = \frac{S_{\bar{y}_x}}{s_y} = 0$ следует, что $S_{\bar{y}_x} = 0$ и, следовательно, $D_{межгр} = 0$. Равенство $D_{межгр} = 0$ означает, что $\bar{y}_x = \bar{y}$, т.е. при всех значениях случайной величины X $\bar{y}_x$ сохраняет постоянное значение, равное $\bar{y}$. Иными словами, при $\eta = 0$ условное среднее $\bar{y}_x$ не является функцией от x, а значит, величина Y не связана корреляционной зависимостью с величиной X. Верно и обратное утверждение: если $\bar{y}_x = const$, т.е. $\bar{y}_{x1} = \bar{y}_{x2} = ... = \bar{y}$, то $D_{межгр} = 0$ и, следовательно, $S_{\bar{y}_x} = 0$, $\eta = 0$.

Свойство 3. Если $\eta = 1$, то Y и X связаны функционально.

Доказательство

Из $\eta = 1$ следует, что

$$S_{\bar{y}_x} = S_y, D_{межгр} = D_{общ} \qquad (1)$$

Так как $D_{общ} = D_{межгр} + D_{внгр}$, то из (1) вытекает, что $D_{внгр} = 0 \Rightarrow D_j = 0$, так что в каждой группе содержатся равные значения y_{ij}, т.е. каждому значению x соответствует одно значение y. Поэтому величины Y и X связаны функционально. Верно и обратное утверждение: если $\bar{y}_{x_i} = y_i$, то и $S_{\bar{y}_x} = S_y$, $\eta = 1$.

Свойство 4. Всегда корреляционное отношение не меньше коэффициента корреляции $\eta \geq |r|$.

Свойство 5. Если $\eta = |r|$, то имеет место точная линейная зависимость. Другими словами, если $\eta = |r|$, то точки (x_i, y_i) лежат на прямой линии регрессии, найденной способом наименьших квадратов.

Убедимся, что с возрастанием η корреляционная связь становится более тесной. Для этого преобразуем соотношение $D_{общ} = D_{межгр} + D_{внгр}$ следующим образом

$$D_{внгр} = D_{общ} \cdot \left(1 - \frac{D_{межгр}}{D_{общ}}\right) = D_{общ} \cdot \left(1 - \eta^2\right) \qquad (2)$$

Из (2) видно, что при $\eta \to 1$

$$D_{внгр} \to 0 \qquad (3)$$

Из (3) вытекает, что

$$D_j \to 0 \qquad (4)$$

Из (4) следует, что $y_{ij} \to \bar{y}_j$, т.е. при $\eta \to 1$ связь величин Y,X становится более тесной, переходя в функциональную при $\eta = 1$

Поскольку в приведенных рассуждениях не делалось никаких допущений о форме корреляционной связи, то η может служить мерой тесноты корреляционной связи любой формы. В этом состоит преимущество корреляционного отношения перед коэффициентом корреляции, который оценивает тесноту лишь линейной связи.

Недостаток: η не позволяет судить, насколько близко расположены точки, найденные по данным наблюдений, к кривой определенного вида, например, к параболе, гиперболе и т.д.

§4.3. Нелинейная корреляционная зависимость

Если график функций регрессии $f(x), \varphi(y)$ изображается кривой линией, то корреляцию называют криволинейной. Например, функции регрессии Y на X могут иметь вид:

$\bar{y}_x = ax^2 + bx + c$ (параболическая корреляция второго порядка)

$\bar{y}_x = ax^3 + bx^2 + cx + d$ (параболическая корреляция 3 порядка)

$\bar{y}_x = \frac{a}{x} + b$ (гиперболическая корреляция)

Теория криволинейной корреляции решает те же задачи, что и теория линейной корреляции:

1) установление формы корреляционной связи;
2) установление тесноты корреляционной связи.

Неизвестные параметры уравнения регрессии ищут методом наименьших квадратов. Для оценки тесноты криволинейной корреляции служит выборочное корреляционное отношение.

Чтобы выяснить суть дела, ограничимся параболической корреляцией 2 порядка, предположив, что данные n наблюдений позволяют считать, что имеет место именно такая корреляция. В этом случае выборочное уравнение регрессии Y на X имеет вид:

$$\bar{y}_x = Ax^2 + Bx + C, \qquad (1)$$

где A,B,C – неизвестные параметры, подлежащие определению.

Пользуясь методом наименьших квадратов, нетрудно получить систему линейных уравнений относительно этих параметров:

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

Найденные из системы (2) параметры A,B,C подставляют в (1) и в итоге получают искомое уравнение регрессии.

Пример. По данным корреляционной таблицы найдем выборочное уравнение регрессии Y на X вида $\bar{y}_x = Ax^2 + Bx + C$:

X / Y	1	1.1	1.2	n_y
6	8	9		17
7		23		23
7.5		1	9	10
n_x	8	33	9	n=50
$\bar{y}_x$	$\frac{6\cdot 8}{8} = 6$	6.74	7.5	

Составим расчетную таблицу

x	n_x	$\bar{y}_x$	$n_x x$	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \bar{y}_x$	$n_x \bar{y}_x x$	$n_x \bar{y}_x x^2$
1.0	8	6	8	8	8	8	48	48	48
1.1	33	6.74	36.3	39.93	43.93	48.32	222.5	244.66	269.13
1.2	9	7.5	10.8	12.96	15.55	18.66	67.50	81	97.20
$\sum$	50		55.1	60.89	67.48	74.98	338	373.66	414.33

Подставив числа нижней строки этой таблицы в (2), получим систему

$$\begin{cases} 74.98A + 67.48B + 60.89C = 414.33 \\ 67.48A + 60.89B + 55.10C = 373.66 \\ 60.89A + 55.10B + 50C = 338 \end{cases}$$

Решив эту систему, найдем $A = 1.95, B = 2.98, C = 1.10$, так что искомое уравнение регрессии имеет вид

$$\bar{y}_x = 1.95\cdot x^2 + 2.98\cdot x + 1.10 \quad (3)$$

При x=1 по исходной таблице $\bar{y}_1 = 6$, а по уравнению (3) $\bar{y}_1 = 6.03$.

§4.4. Множественная корреляционная зависимость

Ранее мы рассматривали корреляционную связь между 2 величинами. Если исследуется связь между несколькими величинами, то корреляцию называют множественной. Рассмотрим случай, когда число величин равно 3 и связь между ними линейная

$$(z-\bar{z})=A(x-\bar{x})+B(y-\bar{y})$$

В этом случае возникают задачи:

1) найти коэффициенты регрессии A,B;
2) оценить силу связи между величиной Z и обоими величинами X,Y;
3) оценить силу связи между Z и X, Z и Y.

Первая задача решается методом наименьших квадратов:

$$A=\frac{\sigma_z}{\sigma_x}\frac{r_{xz}-r_{yz}r_{xy}}{1-r_{xy}^2},\ B=\frac{\sigma_x}{\sigma_y}\frac{r_{yz}-r_{xz}r_{xy}}{1-r_{xy}^2},$$

где r_{xz} - коэффициент корреляции между X и Z; r_{yz} - Y и Z; r_{xy} - X и Y; $\sigma_x,\sigma_y,\sigma_z$ - выборочные средние квадратичные отклонения. Сила связи величины Z с величинами X,Y оценивается выборочным совокупным коэффициентом корреляции

$$R=\sqrt{\frac{r_{xz}^2+r_{yz}^2-2r_{xy}r_{xz}r_{yz}}{1-r_{xy}^2}}$$

Сила связи между Z и X оценивается частным выборочным коэффициентом корреляции

$$r_{xz(y)}=\frac{r_{xz}-r_{xy}r_{yz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{yz}^2\right)}}$$

Сила же связи между Z и Y оценивается частным выборочным коэффициентом корреляции

$$r_{yz(x)}=\frac{r_{yz}-r_{xy}r_{xz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{xz}^2\right)}}$$

Эти коэффициенты имеют те же свойства и тот же смысл, что и обыкновенный выборочный коэффициент корреляции, т.е. служат для оценки линейной связи между величинами.

§4.5. Сглаживание

Часто экспериментальные данные представляют собой зависимость величины f от некоторой другой величины x. Измеренные с некоторой погрешностью или зашумленные экспериментальные данные перед их анализом обычно **сглаживают**. Если количество экспериментальных точек велико, то подбор эмпирической формулы может оказаться весьма затруднительным: формулы с малым числом параметров могут давать большие искажения, а большое число параметров неудобно для анализа. С другой стороны, многие задачи анализа (например, связанные с дифференцированием или интегрированием) не требуют единой аналитической формулы для всех данных. Для анализа важно лишь устранить «шум» эксперимента, сохранив информацию об истинной функции.

Для этой цели применяется сглаживание эмпирических данных, т.е. замена данной таблицы опытных точек другой таблицей близких к ним точек, лежащих на достаточно гладкой кривой.

Сглаживание производится с помощью многочленов (желательно оптимальной степени), приближающих по методу наименьших квадратов выбранные группы опытных точек. Так как наилучшее сглаживание получается для средних точек (когда учитывается информация о поведении функции по обе стороны от сглаживаемой точки), то количество точек для сглаживания выбирают нечетным, а группы точек–скользящими вдоль всей таблицы: берут, например, первые пять точек y_1, y_2, y_3, y_4, y_5 и сглаживают с их помощью среднюю точку y_3, затем берут следующую группу точек y_2, y_3, y_4, y_5, y_6 и сглаживают точку y_4, и т.д. до конца таблицы. При этом остаются несколько крайних точек, которые сглаживаются с меньшей точностью.

Ниже приводятся наиболее употребляемые из простых формул сглаживания для таблиц с постоянным шагом. Сглаженные значения обозначаются волнистой чертой сверху. Основной формулой служит формула сглаживания средней точки, т.е. формула для $\widetilde{f}_i$, остальные формулы применяются только на краях таблицы.

Наиболее простым методом является метод **линейного** сглаживания по **трем** точкам: Линейным сглаживанием называется сглаживание многочленом первой степени.

$$\widetilde{f}_i = \left[f_{i-1} + f_i + f_{i+1}\right]/3,\ i = 1,2,...n-1,$$

$$\widetilde{f}_0 = [5f_0 + 2f_i - f_2]/6,\ i = 0,$$
$$\widetilde{f}_n = [5f_n + 2f_{n-1} - f_{n-2}]/6,\ i = n,$$

где n-номер последней точки, в которой измерена величина f_i.

Метод **линейного** сглаживания по **пяти** точкам основан на использовании формул

$$\widetilde{f}_0 = [3f_0 + 2f_i + f_2 - f_4]/5,\ i = 0,$$
$$\widetilde{f}_1 = [4f_0 + 3f_i + 2f_2 + f_3]/10,\ i = 1,$$
$$\widetilde{f}_i = [f_{i-2} + f_{i-1} + f_i + f_{i+1} + f_{i+2}]/5,\ i = 2,3,...n-2$$
$$\widetilde{f}_{n-1} = [4f_n + 3f_{n-1} + 2f_{n-2} + f_{n-3}]/10,\ i = n-1$$
$$\widetilde{f}_n = [3f_n + 2f_{n-1} + f_{n-2} - f_{n-4}]/5,\ i = n$$

Метод **нелинейного** сглаживания по **семи** точкам обеспечивает усреднение на основе применения полинома третьей степени и реализуется следующими формулами:

$$\widetilde{f}_0 = [39f_0 + 8f_1 - 4(f_2 + f_3 - f_4) + f_5 - 2f_6]/42,$$
$$\widetilde{f}_1 = [8f_0 + 19f_1 + 16f_2 + 6f_3 - 4f_4 - 7f_5 + 4f_6]/42,$$
$$\widetilde{f}_2 = [-4f_0 + 16f_1 + 19f_2 + 12f_3 + 2f_4 - 4f_5 + f_6]/42,$$
$$\widetilde{f}_i = [7f_i + 6(f_{i+1} + f_{i-1}) + 3(f_{i+2} + f_{i-2}) - 2(f_{i+3} + f_{i-3})]/21,$$
$$\widetilde{f}_{n-2} = [-4f_n + 16f_{n-1} + 19f_{n-2} + 12f_{n-3} + 2f_{n-4} - 4f_{n-5} + f_{n-6}]/42$$
$$\widetilde{f}_{n-1} = [8f_n + 19f_{n-1} + 16f_{n-2} + 6f_{n-3} - 4f_{n-4} - 7f_{n-5} + 4f_{n-6}]/42,$$
$$\widetilde{f}_n = [39f_n + 8f_{n-1} - 4(f_{n-2} + f_{n-3} - f_{n-4}) + f_{n-5} - 2f_{n-6}]/42$$

Формулы сглаживания многочленами более высоких степеней не применяются, а формулы сглаживания по большему числу точек применяются крайне редко, так как они оставляют плохо сглаженными большое количество точек по краям таблицы.