Часть III. Элементы корреляционного и регрессионного анализа

Глава 16. Меры связи между случайными величинами

16.1. Предварительные замечания. В ряде случаев при одновременном изучении нескольких признаков, присущих членам статистического коллектива, обнаруживается связь между ними. Желая сделать эту связь наглядной, можно, измерив значения признаков, построить на координатной плоскости точки (x_i, y_i), где $x_i \in X$ *и* $y_i \in Y$ являются значениями первого (X) и второго (Y) признака i-ого члена статистического коллектива. По виду образовавшегося «облака точек» (его называют *корреляционным полем*), можно в какой то мере судить о виде и степени изучаемой связи.

Рассмотрим, например, зависимость веса куска проволоки от его длины при условии, что она имеет одинаковую плотность и одинаковую площадь сечения. Теоретически в этом случае каждому значению длины отрезка будет соответствовать определенное значение веса, а сами точки корреляционного поля расположатся по прямой линии, наклон которой определится плотностью материала и площадью сечения проволоки. Такие зависимости, в которых каждому значению аргумента соответствует одно и только одно значение второй величины, называются функциональными. Как правило, они встречаются в теоретических расчетах.

В реальных условиях, когда значения признаков получают не путем теоретических расчетов, а с помощью измерения, неизбежно возникают ошибки (абсолютно точных измерений просто не бывает), в результате чего равным значениям аргумента (например, длины) может соответствовать не одно, а несколько различных значений второй величины (например, веса).

В таких областях, как биология, педагогика, психология или социология, к ошибкам наблюдения и измерения добавляются, как правило, большие естественные изменения объектов исследования. В опыте с проволокой этому явлению соответствовала бы неоднородность материала и сечения. В результате корреляционное поле, речь о котором шла выше, становится более размытым, а тренд, т.е. изменение, определяющее общее направление развития, выражен менее отчетливо. В этом случае связь между признаками обычно проявляется таким образом, что отдельному значению аргумента соответствует не одно значение второго признака, а целый набор значений с определенным статистическим распределением этих значений. Например, при за-

данном росте мальчики 14 лет имеют целый набор значений веса, составляющих некоторое распределение с определённым средним значением и дисперсией. Такие связи называют *корреляционными связями* или просто *корреляциями.*

С понятием корреляционной зависимости теснейшим образом связано понятие *линии регрессии.* Если каждому значению аргумента $x \in X$ поставить в соответствие среднее арифметическое (математическое ожидание) всех значений $y \in Y$, соответствующих взятому значению x, то корреляционное поле превращается во множество точек, расположенных на некоторой линии, которую называют *линией регрессии*. Если линия регрессии прямая, то регрессию называют *линейной регрессией.*

Если рассматриваемая связь является функциональной, то при любом $x \in X$ вторая случайная величина Y принимает лишь одно определенное значение y и никакого рассеяния точек корреляционного поля около линии регрессии нет.

В случае корреляционной зависимости двух случайных количественных признаков обычным показателем концентрации распределения вблизи линии регрессии служит *корреляционное отношение* $E^2_{yx} = 1 - \sigma^2_{Y|X} / \sigma^2_Y$.

О способе вычисления E^2_{yx} (его часто обозначают символом $\eta^2_{Y|X}$) будет рассказано в пункте 17.3. Там будет показано, что величина E^2_{yx} изменяется от 0 до 1. Нулю она равна тогда и только тогда, когда регрессия имеет вид $\tilde{y}(x) = m_Y$, где m_Y – безусловное математическое ожидание Y. В этом случае говорят, что Y *не коррелирована* с X. Единице величина E^2_{yx} равняется в случае точной функциональной зависимости Y от X.

Как мы видим, интенсивность (сила) корреляционной связи может быть различной – от полной независимости, до функциональной связи. Чем менее размыто корреляционное поле вдоль оси ординат, чем более похоже оно на линию, не имеющую ширину, тем отчетливее, сильнее, интенсивнее корреляционная связь между признаками. По форме корреляция может быть прямолинейной и криволинейной, по направлению – прямой и обратной. Интенсивность корреляционной связи измеряется различными показателями, из которых наиболее распространенным является *коэффициент корреляции* ρ *Бравэ и Пирсона*, о вычислении которого речь пойдет в пункте 16.4.

Однако использование этого коэффициента в качестве меры зависимости оправдано лишь тогда, когда совместное распределение пары (X,Y) нор-

мально или приближенно нормально. Употребление ρ как меры зависимости между произвольными случайными величинами Y и X приводит иногда к ошибочным выводам, так как ρ может равняться нулю даже тогда, когда Y и X связаны строгой функциональной зависимостью (отличной от линейной). Доказывается, что если двумерное распределение X и Y нормально, то линии регрессии Y по X и X по Y являются прямыми линиями. Таким образом, коэффициент корреляции ρ служит мерой зависимости, которой соответствует линейная регрессия.

16.2. Способы задания и изображения корреляционных зависимостей. Задается корреляционная зависимость при помощи корреляционного ряда, состоящего из множества упорядоченных пар значений, из которых первое относится к первому признаку, а второе – ко второму, связанному с первым. В таблице 16.1 приведён фрагмент корреляционного ряда, устанавливающий связь между уровнем JQ и уровнем средней успеваемости учащихся восьмого класса.

Таблица 16.1. Связь между уровнем JQ и средним уровнем успеваемости по математике у школьников восьмого класса

X - уровень JQ	75	80	85	85	90	95	100	100	105	105
Y - ср. успеваемость	3,1	3,3	3,1	3,5	3,5	3,6	3,7	3,8	3,8	4,2

110	110	110	110	115	115	115	120	120	125	125	130	130	135	140
4,0	4,1	4,2	4,4	4,3	4,5	4,6	4,5	4,7	4,6	4,8	4,7	4,9	4,9	5,0

Как было сказано выше, графически корреляционную зависимость можно изобразить в виде корреляционного поля или корреляционной решётки (диаграммы), в которой каждый объект исследования (например, i-ый ученик 8-го класса) отмечается точкой с координатами, соответственно равными $x_i \in X$ и $y_i \in Y$.

При большом объёме статистического коллектива (n) переходят к интервальному распределению случайных величин X и Y. Так, например, в рассмотренном выше примере, значения случайной величины X можно распределить по интервалам: (71; 90), (91; 110), (111; 130), (131; 150). В свою очередь, значения случайной величины Y можно распределить по интервалам: (3,1; 3,5), (3,6; 4,0), (4,1; 4,5), (4,6; 5,0). На этой основе строится корреляционная решётка (табл. 16.2), после чего подсчитывается количество элементов, которые попадают в каждую из образовавшихся клеток.

Таблица 16.2. Корреляционная таблица

Y – Средний балл	X – уровень JQ				
	71-90	91-110…	111-130	131-150	Итого
4,6-5,0	0	0	6	2	8
4,1-4,5	0	4	3	0	7
3,6-4,0	0	5	0	0	5
3,1-3,5	5	0	0…	0	5
Итого	5	9	9	2	25

16.3. Коэффициент корреляции Пирсона. Коэффициент корреляции Пирсона как мера интенсивности (силы) и направления корреляционной связи применяется к зависимостям между двумя признаками, если есть основание считать, что двумерное распределение X и Y нормально, а регрессия линейна, т.е. выражается прямой линией.

При изображении соответствующего корреляционного поля точки, соответствующие объектам наблюдения, заполняют область, напоминающую вытянутый эллипс. Большая ось этого эллипса проходит по диагонали: или от угла наименьших значений (при положительной корреляционной связи), или от угла, где сходятся наименьшие значения одного признака и наибольшие значения другого (при отрицательной корреляционной связи). Соответствующие ситуации изображены на рисунке 16.1.

Вычислим средние арифметические $\bar{x}$ и $\bar{y}$ для каждого из двух вариационных рядов, входящих в состав корреляционного ряда, и отметим на корреляционной решётке точку $M(\bar{x},\bar{y})$. Если теперь рассмотреть произведение $(x_i-\bar{x})(y_i-\bar{y})$, где x_i, y_i – координаты произвольной точки коллектива, то его знак будет зависеть от того, в какой из четвертей, образуемых прямыми $x=\bar{x}$ и $y=\bar{y}$, лежит эта точка. В случае прямой зависимости число точек $M_i(x_i,y_i)$, для которых это произведение будет положительным, будет больше числа точек, для которых оно будет отрицательным. К тому же и абсолютные значения положительных произведений в среднем будут больше абсолютных значений отрицательных произведений (рис. 16.1(а)), поэтому сумма $\overline{y^2}=17{,}5656$, подсчитанная для всех точек коллектива, будет положительна. В случае же обратной связи (рис. 16.1(б)) она будет отрицательной. Если случайные величины (признаки) не имеют систематической связи, то в сумме $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ положительные и отрицательные слагаемые будут частично погашаться, а сама сумма станет близкой к нулю.

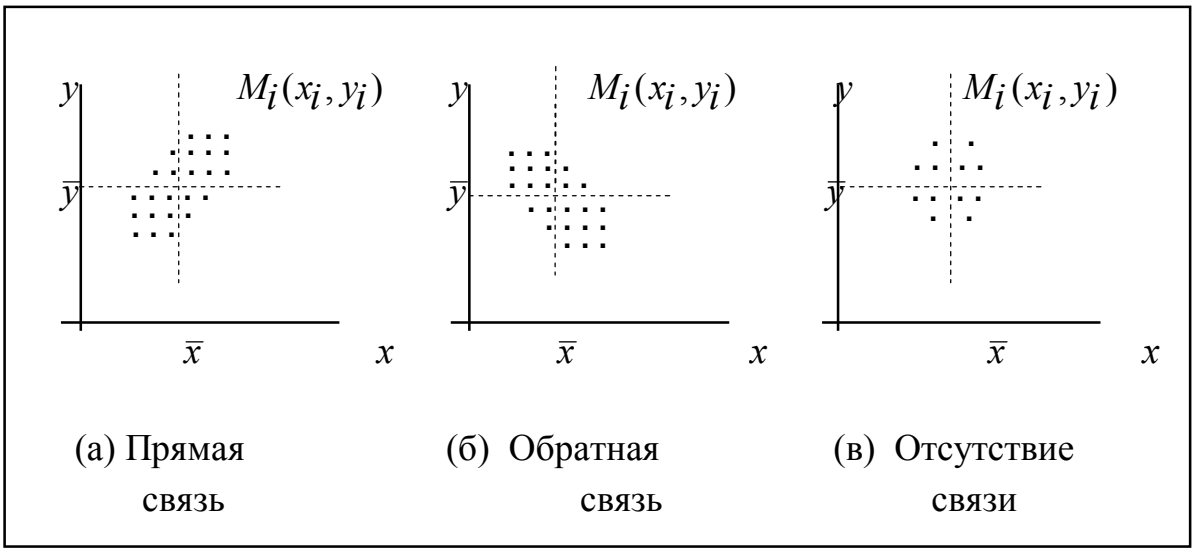


Рис. 16.1. Диаграммы корреляционных полей

Почти очевидно, что на величину рассматриваемой суммы $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$, помимо интенсивности связи между признаками X и Y, которая графически выражается вытянутостью эллипса вдоль главной оси, влияет численность коллектива (n): чем больше n, тем больше слагаемых вида $(x_i-\bar{x})(y_i-\bar{y})$ входит в рассматриваемую сумму. Для того чтобы избавиться от влияния численности коллектива, естественно сумму $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ разделить на n. Однако, как и в случае с дисперсией, её делят на число степеней свободы $n-1$. Величину $S_{xy}=\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ называют *ковариацией* признаков X и Y.

На значение ковариации влияет не только интенсивность связи, но и степень разброса значений x_i и y_i около их средних $\bar{x}$ и $\bar{y}$. Чем больше этот разброс, тем больше будут абсолютные величины произведений $(x_i-\bar{x})(y_i-\bar{y})$. Чтобы избавить меру связи от влияния этого разброса, достаточно ковариацию S_{xy} разделить на стандартные отклонения S_x и S_y. В результате получим меру связи X и Y, не зависящую ни от числа испытаний, ни от степени разброса значений X и Y около $\bar{x}$ и $\bar{y}$. Её называют *коэффициентом корреляции Пирсона* или *произведением моментов.* Если коэф-

фициент корреляции Пирсона подсчитан по всем элементам генеральной совокупности, то его обозначают буквой ρ_{xy}. Таким образом

$$\widehat{\rho}_{xy} = \frac{S_{xy}}{S_x \cdot S_y}. \tag{16.1}$$

Если указанное правило применяется к выборочному корреляционному ряду, то полученное значение называют выборочным коэффициентом корреляции и обозначают символом r_{xy}. Проблемы оценивания ρ_{xy} с помощью выборочного коэффициента корреляции будут рассмотрены в пункте 18.1.

16.4. Вычисление коэффициента корреляции Пирсона. Подставив значения S_{xy}, S_x и S_y и умножив числитель и знаменатель на $n-1$, получим:

$$r_{x\,y} = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2} \cdot \sqrt{\sum \left(y_i - \bar{y}\right)^2}}, \tag{16.2}$$

где суммирование проводится по всем элементам совокупности от 1 до n.

Пользоваться этой формулой при больших значениях n неудобно. Поэтому ее обычно преобразуют к виду, более удобному для расчетов. С этой целью раскроем скобки и просуммировав слагаемые, получим, что

$$r_{xy} = \frac{\sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + \sum \bar{x} \cdot \bar{y}}{\sqrt{\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2} \cdot \sqrt{\sum y_i^2 - 2\bar{y}\sum y_i + n\bar{y}^2}}.$$

Учитывая, что $\sum x_i = n \cdot \bar{x}$, $\sum y_i = n \cdot \bar{y}$, $\sum \bar{x}^2 = n \cdot \bar{x}^2$, $\sum \bar{y}^2 = n \cdot \bar{y}^2$, будем иметь $r_{xy} = \dfrac{\sum x_i y_i - n \cdot \bar{x} \cdot \bar{y} - n \cdot \bar{y} \cdot \bar{x} + n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} \cdot \sqrt{\sum y_i^2 - 2n\bar{y}^2 + n\bar{y}^2}}$, откуда после приведения подобных членов, получим формулу:

$$r_{xy} = \frac{\sum x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \cdot \sqrt{\sum y_i^2 - n\bar{y}^2}}. \tag{16.3}$$

Разделив числитель и знаменатель на n и, введя обозначения

$\dfrac{\sum x_i y_i}{n} = \overline{xy}$, $\dfrac{\sum x_i^2}{n} = \overline{x^2}$; $\dfrac{\sum y_i^2}{n} = \overline{y^2}$, получим окончательно

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}. \qquad (16.4)$$

В качестве примера рассмотрим заданную таблицей 16.3 корреляционную связь между признаками X и Y и подсчитаем соответствующее значение коэффициента корреляции Пирсона.

Таблица 16.3. Корреляционная связь между признаками X и Y

X	1,5	2,5	3,5	3,5	4,0	5,0	5,0	5,5	6,5	6,5	7,5	9,0
Y	1,0	0,8	0,6	1,7	1,7	1,1	3,0	2,0	2,0	3,6	3,0	3,5

Для вычисления r_{xy} «вручную» заполним расчетную таблицу 16.4, построенную в соответствии с формулой 16.4.

Замечание. При использовании компьютера проще всего обратиться к программе Excel. Для этого, занеся в два столбца соответствующие данные и обратившись к «мастер функции» f_n, вызовите раздел «статистика», а в нем функцию «коррел».

Таблица 16.4. Расчет данных, необходимых для вычисления r_{xy}

№	x_i	y_i	x_iy_i	x_i^2	y_i^2		№	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	1,5	1,0	1,5	2,25	1,0		8	5,5	2,0	11,0	30,25	4,0
2	2,5	0,8	2,0	6,25	0,64		9	6,5	2,0	13,0	42,25	4,0
3	3,5	0,6	2,1	12,25	0,36		10	6,5	3,6	23,4	42,25	12,96
4	3,5	1,7	5,95	12,25	2,89		11	7,5	3,0	22,5	56,25	9,0
5	4,0	1,7	6,8	16,0	2,89		12	9,0	3,5	31,5	81,0	12,25
6	5,0	1,1	5,5	25,0	1,21		$\sum$	60,0	24,0	140,25	351,0	60,2
7	5,0	3,0	15,0	25,0	9,0		Сред	5,0	2,0	11,69	29,25	5,02

Разделив суммарные значения на $n = 12$, получим $\bar{x} = 5{,}0$; $\bar{y} = 2{,}0$;

$\overline{xy} = 11{,}69$; $\overline{x^2} = 29{,}25$; $\overline{y^2} = 5{,}02$, откуда, $\hat{r}_{xy} = \dfrac{11{,}69 - 5{,}0 \cdot 2{,}0}{\sqrt{29{,}25 - 25} \cdot \sqrt{5{,}02 - 4}} = 0{,}81.$

16.5. Свойства коэффициента корреляции Пирсона. Коэффициент корреляции Пирсона обладает рядом свойств, из которых приведем два.

Свойство 1. $r\big((ax+b),(cy+d)\big) = \dfrac{ac}{|ac|} \cdot r_{xy}$.

Доказательство. Известно, что $S_{ax+b} = |a| \cdot S_x$, $S_{cy+d} = |c| \cdot S_y$.

$$S_{(ax+b),(cy+d)} = \frac{1}{n-1} \cdot \sum \left[(ax_i + b) - (a\bar{x} + b) \right] \cdot \left[(cy_i + d) - (c\bar{y} + d) \right] =$$

$$= \frac{1}{n-1} \cdot \sum a(x_i - \bar{x}) \cdot c(y_i - \bar{y}) = ac \frac{1}{n-1} \cdot \sum (x_i - \bar{x})(y_i - \bar{y}) = acS_{xy} .$$

$$r_{(ax+b),(cy+d)} = \frac{S_{(ax+b)(cy+d)}}{S_{ax+b} \cdot S_{cy+d}} = \frac{acS_{xy}}{|a|S_x \cdot |c|S_y} = \frac{ac}{|a| \cdot |c|} \cdot \frac{S_{xy}}{S_x \cdot S_y} = \frac{ac}{|ac|} \cdot r_{xy} .$$

Замечание. Используя доказанное свойство, можно упрощать вычисление коэффициента корреляции. В частности, подвергнув элементы первой строки таблицы 16.3 преобразованию $x' = x - 4{,}5$, а элементы второй строки преобразованию $y' = 10y - 20$, получим более удобные для вычисления коэффициента корреляции данные.

Свойство 2. Для любых X,Y коэффициент корреляции Пирсона r_{xy} удовлетворяет неравенству $-1 \le r_{xy} \le 1$.

Доказательство. Рассмотрим ряды $x_1, x_2, ..., x_n$ и $y_1, y_2, ..., y_n$ со статистиками $\bar{x}$, $\overline{x^2}$, S_x и $\bar{y}$, $\overline{y^2}$, S_y. Стандартизируя их, получим ряды: z_{x_1}, z_{x_2}, $\ldots z_{x_n}$ и z_{y_1}, z_{y_2}, $\ldots, z_{y_n}$, в которых $\bar{z}_x = \bar{z}_y = 0$, $S_{z_x} = S_{z_y} = 1$, а $r_{xy} = r_{z_x z_y} = \frac{\sum z_{x_i} z_{y_i}}{\sqrt{\sum z_{x_i}^2} \cdot \sqrt{\sum z_{y_i}^2}}$. Так как $\frac{\sum z_{x_i}^2}{n-1} = S_{z_x} = 1$, то $\sum z_{x_i}^2 = n - 1$. Аналогично $\sum z_{y_i}^2 = n - 1$. Значит, $r_{xy} = \frac{\sum z_{x_i} z_{y_i}}{n-1}$. Оценим теперь $\sum x_i y_i$ для чего рассмотрим очевидное неравенство: $(z_x - z_y)^2 \ge 0$, из которого следует, что $z_x z_y \le \frac{1}{2}\left(z_x^2 + z_y^2\right)$. Тогда $\sum z_{x_i} z_{y_i} \le \frac{1}{2}(\sum z_{x_i}^2 + \sum z_{y_i}^2) = \frac{1}{2}(n-1+n-1) = n-1$.

Теперь можно оценить r_{xy} сверху: $r_{xy} = \frac{\sum z_{x_i} z_{y_i}}{n-1} \le \frac{n-1}{n-1} = 1$. Аналогично можно показать, что $-1 \le r_{xy}$. Откуда $-1 \le r_{xy} \le 1$.

16.6. Вычисление коэффициента корреляции при большом объеме выборки. При большом объёме статистического коллектива (n) от дискретных вариационных рядов, входящих в корреляционный ряд, можно перейти к центрированным интервальным рядам и построить корреляционную таблицу (табл. 16.2). Центрируя интервалы, получим таблицу 16.5.

Таблица 16.5. Центрированно-интервальная корреляционная таблица

		X – уровень JQ				Итого
Y – средний балл		71-90	91-110…	111-130	131-150	
центры		80	100	120	140	
4,6-5,0	4,75	0	0	6	2	8
4,1-4,5	4,25	0	4	3	0	7
3,6-4,0	3,75	0	5	0	0	5
3,1-3,5	3,25	5	0	0	0	5
Итого		5	9	9	2	25

Из таблицы видно, что теперь вместо двадцати пяти пар значений из X и Y мы имеем всего шесть пар, большинство из которых повторяются, например, пара (80; 3,25) – 5 раз, пара (100; 3,75) – 5 раз, пара (100; 4,25) – 4 раза, и т.д. Для упрощения расчетов подвергнем значения случайной величины X-преобразованию $a_x = (x - 100):20$, а значения случайной величины Y-преобразованию $a_y = (y - 3{,}75):0{,}5$. В результате получим таблицу 16,6.

Таблица 16.6. Корреляционная таблица после преобразования переменных

Переменная a_y	Переменная a_x				Итого
	1	0	1	2	
+2			6	2	8
+1		4	3		7
0		5			5
-1	5				5
Итого	5	9	9	2	25

Расчет коэффициента корреляции приведен в таблице 16.7.

Таблица 16.7. Расчётная корреляционная таблица

a_y	Значения a_x				n_y	$n_y \cdot a_y$	$n_y \cdot a_y^2$
	-1	0	+1	+2			
+ 2			6	2	8	16	32
+ 1		4	3		7	7	7
0		5			5	0	0
- 1	5				5	- 5	5
n_x	5	9	9	2	25	$\sum n_y a_y = 18$	$\sum n_y a_y^2 = 44$
$n_x \cdot a_x$	- 5	0	9	4	$\sum n_x a_x = 8$		
$n_x \cdot a_x^2$	5	0	9	8	$\sum n_x a_x^2 = 22$		

Из данных, приведённых в таблице 16.7, получаем:

$$\bar{x} = \frac{\sum n_x \cdot a_x}{n} = \frac{8}{25} = 0{,}32\ ; \qquad \bar{y} = \frac{\sum n_y \cdot a_y}{n} = \frac{18}{25} = 0{,}72;$$

$$\overline{x^2} = \frac{\sum n_x \cdot a_x^2}{n} = \frac{22}{25} = 0{,}88\ ; \qquad \overline{y^2} = \frac{\sum n_y \cdot a_y^2}{n} = \frac{44}{25} = 1{,}76\ ;$$

$$\overline{xy} = \left(5 \cdot (-1) \cdot (-1) + 5 \cdot 0 \cdot 0 + 4 \cdot 0 \cdot 1 + 3 \cdot 1 \cdot 1 + 6 \cdot 1 \cdot 2 + 2 \cdot 2 \cdot 2\right) : 25 = 1{,}12;$$

$$r_{xy} = \frac{1{,}12 - 0{,}32 \cdot 0{,}72}{\sqrt{0{,}88 - 0{,}32^2} \cdot \sqrt{1{,}76 - 0{,}72^2}} \approx \frac{0{,}8896}{0{,}8818 \cdot 1{,}1143} \approx 0{,}905.$$

Если бы все расчёты велись по 25 исходным парам, то, как можно проверить $\bar{x} = 109{,}2$; $\bar{y} = 4{,}152$; $\overline{x^2} = 12216$; $\overline{y^2} = 17{,}5656$; $\overline{xy} = 462{,}84$, откуда $r_{xy} = 0{,}968$. Потеря точности – цена за упрощение расчётов.

16.7. Интерпретация значений коэффициента корреляции Пирсона. Наличие корреляционной связи двух переменных отнюдь не означает, что между ними существует причинно-следственная связь. Взаимосвязи переменных в педагогике и общественных науках почти всегда слишком сложны, чтобы их объяснением могла служить единственная причина. К тому же, часто наблюдаемая связь существует благодаря другим переменным. Наконец, даже если можно предположить причинную связь, коэффициент r_{xy} ничего не говорит о её направленности, что является причиной, а что - следствием. Однако хотя корреляция прямо не указывает на причинную связь, она может служить ключом к разгадке причин. Иногда отсутствие корреляции может оказывать более глубокое воздействие на нашу гипотезу о причинной связи, чем наличие сильной корреляции. Сильная корреляционная связь – это факт, который в разных ситуациях можно объяснить по-разному:

1) наличием причинно - следственной связи;
2) наличием общей причины, влияющей и на X и на Y;
3) объединением двух групп, в каждой из которых X и Y связи не имеют.

Последний случай рассмотрим подробнее. Пусть изучается связь между успеваемостью школьников и ощущением тревоги, которое они испытывают при написании контрольной работы. На рисунке 16.2 изображена корреляционная решётка, на которой буквами М и Д обозначены члены коллектива: М – мальчики, Д – девочки. Если не различать данные по мальчикам и девочкам, то мы получим значение $r_{xy} > 0$. Возникает оно потому, что у мальчиков и успеваемость, и тревожность оказались ниже, чем у девочек. В пределах же каждой из этих групп коэффициент корреляции равен нулю.

Из всех способов, которыми могут быть связаны две переменные, r_{xy} оценивает только один. Величина r_{xy} представляет собой меру интенсивности линейной связи X и Y. Если X и Y имеют некоторую нелинейную связь, то близкие к нулю значения r_{xy} могут быть получены, даже несмотря на то, что X и Y сильно связаны (рис. 16.3).

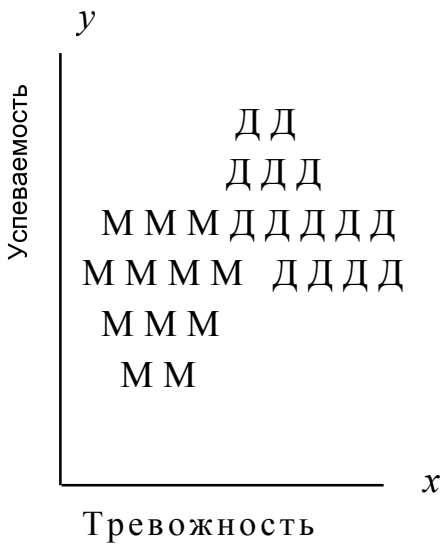


Рис. 16.2. Связь между успеваемостью и тревожностью

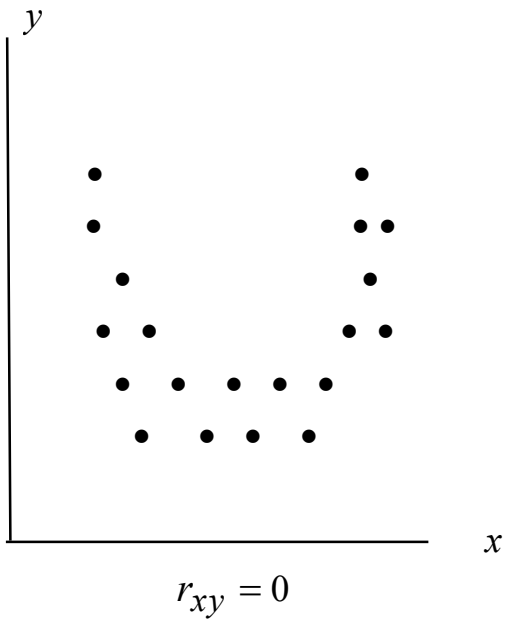


Рис. 16.3. Нулевая корреляция при сильной причинно-следственной связи

Часто оценки педагогических и психологических тестов дают «потолочные» или «подвальные» эффекты. Они возникают тогда, когда испытания слишком лёгкие и подавляющее большинство учащихся получает самые высокие оценки, или слишком трудные – и большинство учащихся получает самые низкие оценки.

Интерпретация r_{xy} зависит от формы распределений X и Y и их совместного распределения. Для корректности интерпретации r_{xy} необходимо, чтобы они были распределены нормально. При других распределениях пределы изменения r_{xy} могут сужаться. Если x сильно скошено положительно, а y – отрицательно, то максимальное значение r_{xy} может быть уже не 1, а, например, 0,60.

Глава 17. Элементы регрессионного анализа

17.1. Уравнения линейной регрессии. Вернёмся к вопросу о линиях регрессии. Допустим, что корреляционная зависимость Y от X близка к функциональной. В этом случае можно ставить вопрос о функции, которая в *среднем* выражает зависимость Y от X. Наиболее естественно в этом случае взять функцию, принимающую при каждом фиксированном x значение, равное среднему арифметическому соответствующих ему значений величины Y (при данном x). Её обозначают $\tilde{y} = \varphi(x)$. Волнистый знак над y, его называют «тильда», означает то обстоятельство, что в качестве образа x берётся среднее арифметическое из соответствующих ему значений y. Линия на координатной плоскости (X, Y), определяемая уравнением $\tilde{y} = \varphi(x)$, называется линией регрессии Y на X.

Аналогично определяется линия регрессии X на Y; её уравнение есть $\tilde{x} = \psi(y)$, в котором каждому значению y из Y ставится в соответствие среднее арифметическое всех тех значений x, которые соответствуют взятому значению y.

В статистике часто приходится рассматривать такие системы (X, Y), для которых обе линии регрессии представлены прямыми, например, $\tilde{y} = ax + b$ и $\tilde{x} = cy + d$. В этом случае говорят о линейной корреляции между X и Y.

Предположим, что у нас есть основание считать, что корреляционная связь между признаками X и Y близка к линейной. Тогда возникает проблема построения прямой $y = ax + b$, наименее уклоняющейся от точек корреляционного поля.

В качестве основного критерия «наименьшего уклонения» принимается требование, чтобы сумма квадратов отклонений точек корреляционного поля от искомой прямой (вдоль оси Y) была минимальной. Рассмотрим некоторое корреляционное поле, соответствующее переменным X и Y, и прямую $y = ax + b$. Зафиксируем произвольное значение x_i из X и все соответствующие ему значения из Y. Их может оказаться несколько. Обозначим их y_{i1}, y_{i2},...,y_{ik}. Все точки $(x_i, y_{i1}), (x_i, y_{i2}), \ldots, (x_i, y_{ik})$ лежат на одной вертикальной прямой $x = x_i$. Величину $e_{ij} = y_{ij} - (ax_i + b)$ называют *ошибкой оценивания* (ошибкой предсказания). На рисунке 17.1 она изображена в виде вертикального отрезка.

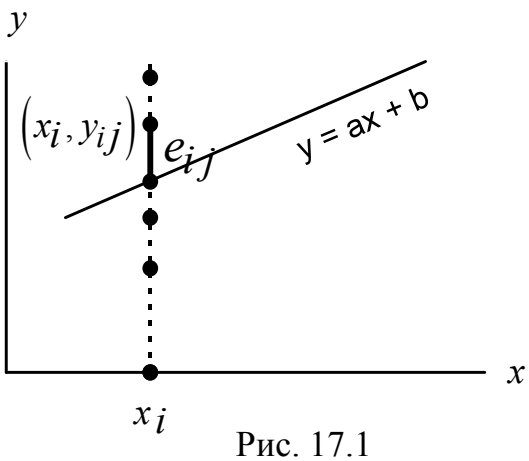


Рис. 17.1

Таким образом, для того чтобы прямая $y = ax + b$ наименее уклонялась от точек заданного корреляционного поля, необходимо, чтобы сумма $\sum e_{ij}^2$, в которой суммирование ведется по всем точкам корреляционного поля, была минимальной.

Так как $\sum e_{ij}^2 = \sum\left[y_{ij} - \left(ax_i + b\right)\right]^2$, то проблема сводится к нахождению значений параметров a и b, при которых функция $\Phi(a,b) = \sum\left[y_{ij} - \left(ax_i + b\right)\right]^2$ принимает наименьшее значение.

Как известно, для этого нужно прежде всего найти значения a и b, при которых производные Φ'_a и Φ'_b равны нулю, т.е. решить систему:

$$\Phi'_a = -2\sum\left[y_{ij} - \left(ax_i + b\right)\right]x_i = 0\ ,\ \Phi'_b = -2\sum\left[y_{ij} - \left(ax_i + b\right)\right] = 0\ .$$

Суммирование ведётся по всем n точкам корреляционного поля.

После очевидных преобразований получим систему двух уравнений с двумя неизвестными a и b:

$$\sum x_i y_{ij} - a\sum x_i^2 - b\sum x_i = 0\ ,\quad \sum y_{ij} - a\sum x_i - nb = 0\ .$$

Разделив оба уравнения на n и введя уже привычные для нас обозначения $\left(\overline{xy}, \bar{x}, \bar{y}\right)$, получим систему:

$$\overline{xy} - a\overline{x^2} - b\bar{x} = 0\ ,\quad \bar{y} - a\bar{x} - b = 0\ .$$

Решив её, найдём: $a = \dfrac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}$, $b = \bar{y} - a\bar{x}$.

Таким образом, искомая линейная зависимость y от x имеет вид:

$$y = ax + b\ ,\ \text{где}\ a = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}\ ,\ b = \bar{y} - a\bar{x}\ . \qquad (17.1)$$

Если корреляция между признаками X и Y линейна, то построенная нами прямая (рис 17.1) будет линейной регрессией Y на X: $\tilde{y} = ax + b$. Часто это название применяют к прямой (17.1) и в том случае, если связь X и Y не является строго линейной. Легко проверить, что коэффициент линей-

ной регрессии a связан с коэффициентом корреляции r_{xy} Пирсона следующим соотношением

$$a = r_{xy} \cdot \frac{s_y}{s_x} \ . \qquad (17.2)$$

Докажем это: $a = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}} \cdot \dfrac{\sqrt{\overline{y^2} - \bar{y}^2}}{\sqrt{\overline{x^2} - \bar{x}^2}}$.

Первый множитель равен r_{xy} (см. формулу 16.4), второй, после умножения числителя и знаменателя на $\sqrt{\dfrac{n}{n-1}}$, будет равен $\dfrac{s_y}{s_x}$, откуда и следует равенство (17.2).

17.2. Построение линейной регрессии. В качестве примера рассмотрим корреляционную связь между случайными величинами X и Y.

X	13	15	27	30	25	10	22	27	20	18
Y	17	14	21	25	27	11	19	29	22	12

Расчёт всех необходимых данных для вычисления параметров a и b уравнения линейной регрессии Y по X приведён в таблице 17.1.

Таблица 17.1. Расчёт параметров уравнения линейной регрессии

№ п/п	x_i	y_i	x_i^2	y_i^2	$x_i y_i$	№ п/п	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	13	17	169	289	221	7	22	19	484	361	418
2	15	14	225	196	210	8	27	29	729	841	783
3	27	21	729	441	567	9	20	22	400	484	440
4	30	25	900	625	750	10	18	12	324	144	216
5	25	27	625	729	675	Итого	207	197	4685	4231	4390
6	10	11	100	121	110	Сред.	20,7	19,7	468,5	423,1	439,0

Используя итоговые данные таблицы, получим: $\bar{x}$=20,7; $\bar{y}$=19,7; $\overline{x^2}$=468,5; $\overline{y^2}$=423,1; $\overline{xy}$=439,0, откуда $\hat{a} = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{439 - 20{,}7 \cdot 19{,}7}{468{,}5 - 20{,}7^2} = 0{,}78$, $\hat{b} = \bar{y} - a\bar{x}$=19,7-0,78·20,7=3,55. Поэтому уравнение линейной регрессии Y по X имеет вид $\tilde{y} = 0.78x + 3.55$.

Теперь легко сопоставить экспериментальные значения y_{ij} со значениями, предсказываемыми формулой $\tilde{y} = 0.78x + 3.55$, и вычислить значения ошибок оценивания – e_{ij}. Соответствующие результаты приведены в таблице 17.2.

Таблица 17.2. Фактические и расчётные значения случайной величины Y (ошибки оценивания).

x_i	13	15	27	30	25	10	22	27	20	18
y_i	17	14	21	25	27	11	19	29	22	12
$\hat{y}_i$	13,7	15,3	24,6	27,0	23,1	11,4	20,7	24,6	19,2	17,6
e_i	+3,3	-1,3	-3,6	+2,0	+3,9	-0,4	-1,7	+4,4	+2,8	-5,6
e_i^2	10,89	1,69	12,96	4	15,21	0,16	2,89	19,36	7,84	31,36

При найденных значениях a и b сумма квадратов ошибок оценивания $\sum_{i-1}^{10} e_i^2 = 106{,}36$.

На рисунке 17.2 изображена найденная нами прямая и эмпирические точки $(x_i . y_i)$.

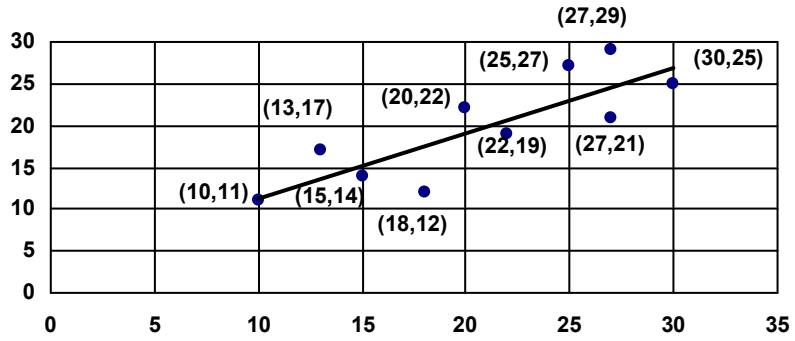


Рис. 17.2

При любой другой прямой эта сумма будет больше. Полученную нами формулу и построенную прямую в определённых условиях можно использовать для «предсказания» наиболее вероятных значений Y, соответствующих значению x. Например, можно ожидать, что при $x = 17$, y будет равен 16,8.

Для того чтобы иметь возможность предсказывать значения X по значениям Y, необходимо вычислить параметры уравнения прямой $\tilde{x} = cy + d$, наименее уклоняющейся вдоль оси OX от точек корреляционного поля.

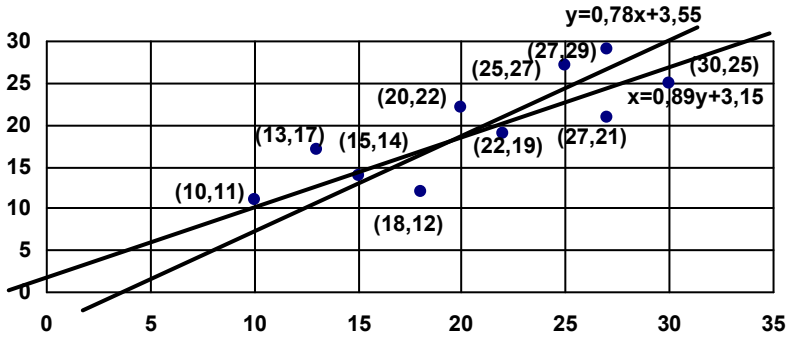


Рис. 17.3

Проведя рассуждения, аналогичные изложенным в пункте 17.1, получим уравнение

$$\tilde{x} = cy + d, \quad \text{где} \quad c = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{y^2} - \bar{y}^2}, \text{ а} \quad d = \bar{x} - c\bar{y}. \tag{17.3}$$

В рассмотренном нами примере $\hat{c} = \dfrac{439 - 20{,}7 \cdot 19{,}7}{423{,}1 - 19{,}7^2} = \dfrac{439 - 407{,}79}{35{,}01} = 0{,}89,$

а $\hat{d} = 20{,}7 - 0{,}89 \cdot 19{,}7 = 3{,}15$, и, следовательно, искомое уравнение, выражающее зависимость X, от Y, будет иметь вид $\tilde{x} = 0{,}89y + 3{,}15$. Обе прямые изображены на рисунке 17.3.

17.3. Измерение нелинейных связей между переменными. Корреляционное отношение. Рассмотренные выше методы предсказания и оценивания степени корреляционной связи между признаками X и Y только тогда приводят к значениям, которые можно надёжно интерпретировать, когда связь между X и Y линейна или хотя бы близка к линейной. В противном случае, т.е. если рассматриваемая связь далека от линейной, в чём можно убедиться, построив корреляционное поле, приходится обращаться к иным методам.

Рассмотрим, например, корреляционное поле, заданное на рисунке 17.4.

Хотя выраженная на нём корреляционная связь и не линейная, но зависимость Y от X в среднем достаточно отчётлива: с увеличением значе-

ний X значения Y в среднем сначала растут, а после достижения наибольшего значения (примерно при $x = 24$) начинают убывать.

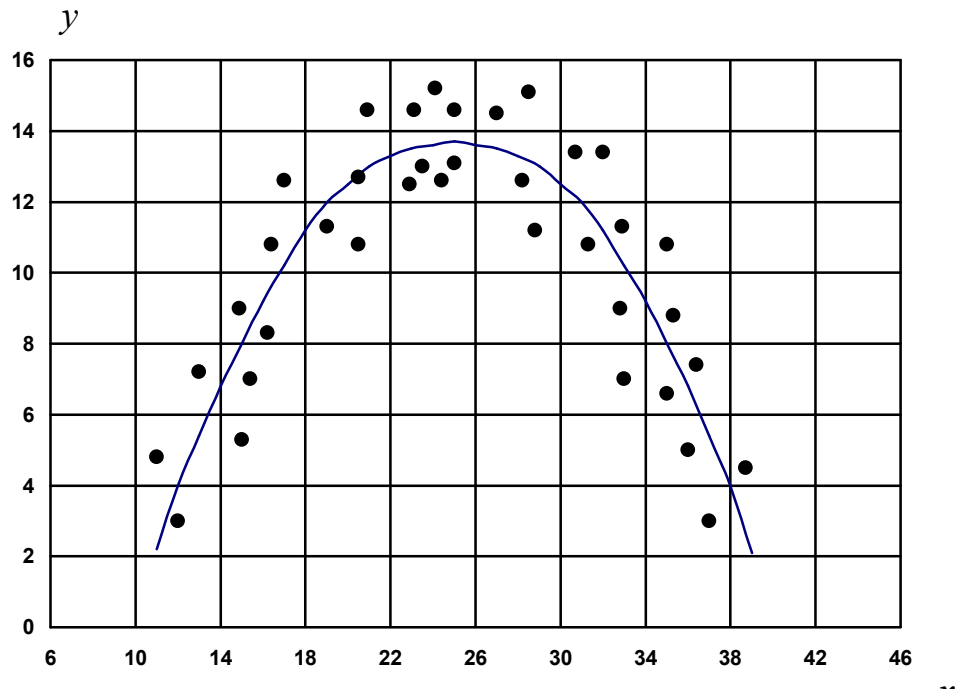


Рис. 17.4

Регрессионная линия $\tilde{y} = \varphi(x)$, по-видимому, близка параболе $\tilde{y} = ax^2 + bx + c$, коэффициенты которой, как и в случае линейной зависимости, можно было бы вычислить, минимизируя сумму квадратов отклонений y_{ij} от вычисленных $\tilde{y}_i$ вдоль оси ординат. О степени концентрации точек корреляционного поля вблизи линии регрессии $\tilde{y} = \varphi(x)$ в общем случае, как для линейной, так и для нелинейной связи, судят по величине так называемого *корреляционного отношения* $\eta^2_{y/x}$ (ню-квадрат Y по X) или, что проще для написания, E^2_{yx}.

Вычисляется корреляционное отношение по формуле:

$$E^2_{yx} = 1 - \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{s} n_{ij}\left(y_{ij} - \hat{y}_i\right)^2}{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{s} n_{ij}\left(y_{ij} - \bar{y}\right)^2}, \qquad (17.4)$$

где k – число столбцов, а s – число строк таблицы 17.3, $\bar{y}$ – среднее арифметическое *всех* значений Y, а $\hat{y}_i$ – среднее арифметическое значений Y,

соответствующих фиксированному значению x_i. Суммирование в обоих случаях ведётся по всем точкам корреляционного поля.

Проиллюстрируем правило вычисления E^2_{yx} на примере приведённого выше (рис. 17.4) корреляционного поля. Для численной обработки данных их необходимо сгруппировать и представить в форме корреляционной таблицы, в каждой клетке которой приводится численность n_{ij} тех пар (x,y), компоненты которых попадают в соответствующие интервалы группировки по каждой переменной.

Предполагая длины интервалов группировки (по каждому из переменных) равными между собой, выбирают центры x_i (соответственно y_i) этих интервалов и числа n_{ij} в качестве основы для расчётов. В нашем случае получится корреляционная таблица 17.3.

Таблица 17.3. Корреляционная таблица

Середина интервала y_{ij}		Середина интервала x_i									
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8		
		12	16	20	24	28	32	36	40	n_{x_i}	$\bar{x}_i$
y_{i1}	15			1	3	2				6	24,7
y_{i2}	13		1	1	4	1	2			9	24,9
y_{i3}	11		1	2		1	2	1		7	26,3
y_{i4}	9		2				1	1		4	25,0
y_{i5}	7	1	1				1	2		5	26,4
y_{i6}	5	1	1					1	1	4	26,0
y_{i7}	3	1						1		2	24,0
n_{y_i}		3	6	4	7	4	6	6	1	37	25,4
y_i		5,0	9,0	12,5	13,9	13,5	10,7	7,0	5,0	10,3	

Из двух сумм $\sum\limits_{i=1}^{л}\sum\limits_{j=1}^{s} n_{ij}\left(y_{ij}-\hat{y}_i\right)^2$ и $\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{s} n_{ij}\left(y_{ij}-\bar{y}\right)^2$ проще вычисляется сумма вторая, которая отличается от дисперсии S_y множителем $n-1$.

$$\sum_{i=1}^{k}\sum_{j=1}^{s} n_{ij}\left(y_{ij}-\bar{y}\right)^2 = 6(15-10{,}3)^2 + 9(13-10{,}3)^2 + 7(11-10{,}3)^2 +$$

$$+ 4(9-10{,}3)^2 + 5(7-10{,}3)^2 + 4(5-10{,}3)^2 + 2(3-10{,}3)^2 = 481{,}73.$$

Для нахождения суммы $\sum_{i=1}^{л}\sum_{j=1}^{s} n_{ij}\left(y_{ij}-\hat{y}_i\right)^2$ вычислим её сначала для каждого столбца таблицы и результаты сложим. В нашем случае первая из этих сумм равна

$$\sum_{i=1}^{7} n_{1j}\left(y_{1j}-\hat{y}_1\right)^2 = 1\cdot\left(y_{15}-5\right)^2 + 1\cdot\left(y_{16}-5\right)^2 + 1\cdot\left(y_{17}-5\right)^2 =$$

$$= (7-5)^2 + (5-5)^2 + (3-5)^2 = 8.$$

Вторая сумма при $i=2$ (по данным второго столбца) $\sum_{j=1}^{7} n_{2j}\left(y_{2j}-\hat{y}_2\right)^2 = 40$. Аналогично третья сумма (по данным третьего столбца) $\sum_{j=1}^{7} n_{3j}\left(y_{3j}-\hat{y}_3\right)^2 = 11$, четвёртая $\sum_{j=1}^{7} n_{4j}\left(y_{4j}-\hat{y}_4\right)^2 = 6,87$; пятая $\sum_{j=1}^{7} n_{5j}\left(y_{5j}-\hat{y}_5\right)^2 = 11$; шестая $\sum_{j=1}^{7} n_{6j}\left(y_{6j}-\hat{y}_6\right)^2 = 27{,}3$; седьмая $\sum_{j=1}^{7} n_{7j}\left(y_{7j}-\hat{y}_7\right)^2 = 40$; восьмая $\sum_{j=1}^{7} n_{8j}\left(y_{8j}-\hat{y}_8\right)^2 = 0$.

Суммируя все эти значения, получим:

$$\sum_{i=1}^{л}\sum_{j=1}^{s} n_{ij}\left(y_{ij}-\hat{y}_i\right)^2 = 144{,}21.$$

В результате значение корреляционного отношения

$$E^2_{yx} = 1 - \frac{\Sigma\left(y_{ij}-\hat{y}_i\right)^2}{\Sigma\left(y_{ij}-\bar{y}\right)^2} = 1 - 144{,}21/481{,}73 = 0{,}71.$$

Если учесть, что

$$\sum_{i=1}^{k}\sum_{j=1}^{s}\left(y_{ij}-\hat{y}_i\right) = \sum_{i=1}^{k}\sum_{j=1}^{s}\left(y_{ij}-\bar{y}\right)^2 + \sum_{i=1}^{k} n_i\left(\bar{y}_i-\hat{y}_i\right)^2,$$

можно получить более удобную для вычисления формулу

$$E_{yx} = \frac{S_1 - R}{S_2 - R},$$

(17.5)

в которой:

$$S_1 = \sum_{i=1}^{k} \Big(\sum_{j=1}^{s} n_{ij} y_j \Big)^2 / n_i, \quad S_2 = \sum_{j=1}^{s} n_{\cdot j} y_j^2, \quad R = \left(\sum_{j=1}^{s} n_{\cdot j} y_j \right)^2 / n.$$

Воспользовавшись этой формулой, получим:

$$S_1 = \frac{15^2}{3} + \frac{54^2}{6} + \frac{50^2}{4} + \frac{97^2}{7} + \frac{54^2}{4} + \frac{64^2}{6} + \frac{42^2}{6} + \frac{5^2}{1} \approx 4261.$$

$$S_2 = 6 \cdot 15^2 + 9 \cdot 13^2 + 7 \cdot 11^2 + 4 \cdot 9^2 + 5 \cdot 7^2 = 4405.$$

$$R = (90 + 117 + 77 + 36 + 35 + 20 + 6)^2 / 37 = 381^2 / 37 \approx 3923.$$

Откуда $\hat{E}^2_{yx} = \dfrac{4261 - 3923}{4405 - 3923} \approx 0{,}70$.

Если бы мы, невзирая на вид корреляционного поля, расположение точек на котором явно не линейно, вычислили коэффициент корреляции Пирсона, то обнаружили бы, что он равен

$$\hat{r}_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}} = \frac{260{,}6 - 25{,}4 \cdot 10{,}3}{\sqrt{709{,}6 - 25{,}4^2} \cdot \sqrt{119{,}1 - 10{,}3^2}} = \frac{-1{,}02}{8{,}03 \cdot 3{,}61} = -0{,}035.$$

Такое близкое к нулю значение коэффициента корреляции Пирсона получилось потому, что корреляционная зависимость Y от X не линейная. В связи с этим ещё раз подчеркнём, что коэффициент корреляции Пирсона измеряет силу *линейной* зависимости Y от X, степень концентрации точек корреляционного поля около прямой линейной регрессии. Если же связь между X и Y не линейная, то пользоваться коэффициентами корреляции Пирсона не имеет смысла. В этом случае пользуются корреляционным отношением E^2_{yx}, значение которого, как мы видели, достаточно велико и свидетельствует о высокой концентрации точек корреляционного поля около кривой регрессии. Корреляционное отношение характеризует степень концентрации точек корреляционного поля около любой линии регрессии, в том числе и около прямой. В последнем случае его значение близко значению квадрата коэффициента корреляции Пирсона.

Основные свойства корреляционного отношения:

1. Корреляционное отношение E^2_{yx} удовлетворяет неравенству $0 \le E^2_{yx} \le 1$.

2. Равенство $E^2_{yx} = 0$ соответствует некоррелированным случайным величинам.

3. $E^2_{yx} = 1$ тогда и только тогда, когда имеется точная функциональная связь между X и Y.

4. В случае линейной зависимости Y от X корреляционное отношение совпадает с квадратом коэффициента корреляции, т.е. $r^2_{xy} = E^2_{yx}$.

5. Корреляционное отношение несимметрично относительно X и Y, поэтому наряду с E^2_{yx} рассматривается E^2_{xy} - корреляционное отношение X по Y, определяемое аналогичным образом. Между E^2_{yx} и E^2_{xy} нет какой-либо простой зависимости.

6. Величина $E^2_{yx} - r^2_{xy}$ используется в качестве индикатора отклонения регрессии от линейной.

Глава 18. Оценка параметров корреляционных связей и регрессий

18.1. Доверительный интервал для коэффициента корреляции генеральной совокупности. Создание оценочных критериев для генеральной корреляции ρ, как и при оценке генеральной средней, можно начать с построения выборочного распределения для коэффициента r. Для этого из заданной генеральной совокупности с двумя признаками последовательно извлекаются выборки объёма n, вычисляются значения выборочного коэффициента корреляции (r) и строится выборочное распределение. В отличие от выборочного распределения для среднего $\bar{x}$, выборочное распределение для коэффициента корреляции r зависит от значения ρ.

Если ρ=0, то выборочное распределение для r будет близким к нормальному с нулевым средним ($M[r]=0$), дисперсией $D[r]=\frac{1}{n-1}$ и стандартным отклонением $\sigma[r]=\frac{1}{\sqrt{n-1}}$, которое называют *стандартной ошибкой* (или ошибкой репрезентативности*) коэффициента* корреляции.

Если же $\rho \neq 0$, то выборочное распределение r уже не является нормальным. С увеличением ρ от 0 до 1 в выборочном распределении r резко возрастает отрицательная асимметрия. С уменьшением ρ от 0 до -1 увеличивается положительная асимметрия. Кроме того, $M[r] \neq \rho$, а стандартное отклонение $\sigma[r]=\sqrt{\frac{1-\rho^2}{n-1}}$ зависит от ρ и меняется с изменением последнего.

В первом случае, когда $\rho=0$, значения коэффициента корреляции, вычисленные для случайных выборок объема n, с надежностью γ должны удовлетворять неравенству $-_{\alpha/2}z \cdot \frac{1}{\sqrt{n-1}} < r_{xy} < {}_{\alpha/2}z \cdot \frac{1}{\sqrt{n-1}}$, где $\alpha=1-\gamma$, и, следовательно, попадать в интервал $(-_{\alpha/2}z\frac{1}{\sqrt{n-1}};\ {}_{\alpha/2}z\frac{1}{\sqrt{n-1}})$. Если вычисленное значение r_{xy} отклонится от нуля на величину, большую ${}_{\alpha/2}z\frac{1}{\sqrt{n-1}}$, то на уровне значимости $\alpha=1-\gamma$ надо отклонить предположение о равенстве нулю коэффициента корреляции генеральной совокуп-

ности. Следовательно, одним из средств проверки нуль-гипотезы о равенстве нулю коэффициента корреляции генеральной совокупности может служить критерий $z = \dfrac{r_{xy}}{1/\sqrt{n-1}}$.

Более корректной в этом случае является статистика:

$$t = \frac{r_{xy}}{\sqrt{\left(1 - r_{xy}^2\right)/\left(n-2\right)}}, \tag{18.1}$$

достаточно точно описываемая t-распределением Стьюдента с числом степеней свободы $\nu = n - 2$.

Во втором случае, т.е. когда $\rho \neq 0$, построение доверительного интервала существенно усложняется. Приходится прибегать к так называемому преобразованию Р.А.Фишера:

$$z_r = \frac{1}{2}\ln\frac{1+r}{1-r}. \tag{18.2}$$

Если каждое выборочное значение r подвергнуть этому преобразованию и построить выборочное распределение не для r, а для z_r, то оно будет приближённо нормальным со средним $M[z_r] = z_\rho$, которое получается из ρ z-преобразованием Фишера и стандартным отклонением

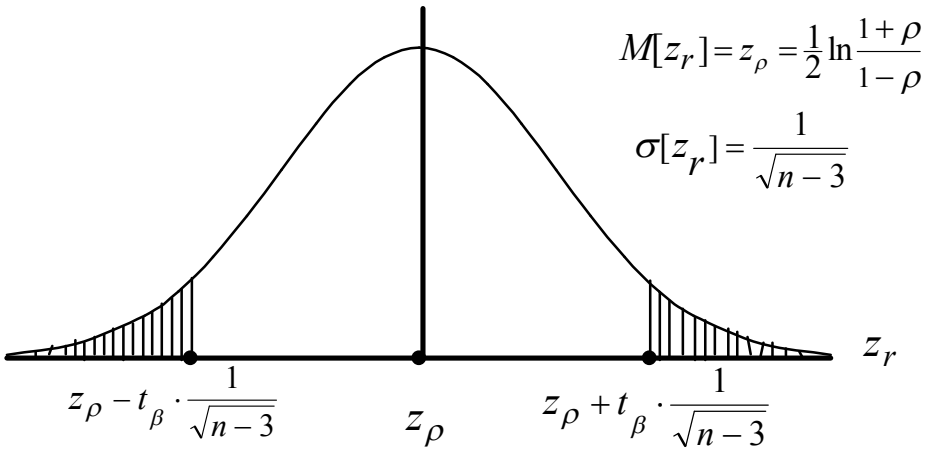


Рис.18.1 Выборочное распределение для z_r

$\sigma[z_r] = \frac{1}{\sqrt{n-3}}$, уже не зависящим от ρ.

Так как полученное распределение z_r нормально (точнее, почти нормально), то интервал $\left(z_\rho - 1{,}96 \cdot \dfrac{1}{\sqrt{n-3}};\quad z_\rho + 1{,}96 \cdot \dfrac{1}{\sqrt{n-3}}\right)$ содержит 95

процентов всех значений z_r, а интервал $\left(z_\rho - 2{,}58\cdot\frac{1}{\sqrt{n-3}};\ z_\rho + 2{,}58\cdot\frac{1}{\sqrt{n-3}}\right)$ – 99 процентов этих значений и т.д.

Преобразовав неравенство $z_\rho - \alpha/2\, t\cdot\frac{1}{\sqrt{n-3}} \le z_r \le z_\rho + \alpha/2\, t\cdot\frac{1}{\sqrt{n-3}}$ в неравенство $z_r - \alpha/2\, t\cdot\frac{1}{\sqrt{n-3}} \le z_\rho \le z_r + \alpha/2\, t\cdot\frac{1}{\sqrt{n-3}}$, получим оценку z_ρ в виде $z_\rho = z_r \pm \alpha/2\, t\cdot\frac{1}{\sqrt{n-3}}$. Теперь, чтобы получить оценку ρ, достаточно выполнить преобразование, обратное z-преобразованию Фишера:

$$r = \frac{e^{2z}-1}{e^{2z}+1}. \qquad (18.3)$$

В качестве примера использования полученных оценок возьмём выборку из 200 наблюдений, извлечённую случайным образом из воображаемой двумерной нормальной совокупности. Пусть r оказалось равным 0,315. Требуется найти доверительный интервал с доверительной вероятностью $\gamma = 0{,}95$.

Для этого, прежде всего, с помощью z-преобразования Фишера (18.2) преобразуем $r = 0{,}315$ в $z_r = 0{,}326$. Затем найдём $\alpha/2\, t\cdot\frac{1}{\sqrt{n-3}}$, которое при $\alpha/2\, t = 1{,}96$ будет равно $1{,}96\cdot\frac{1}{\sqrt{200-3}} = 0{,}140$. Следовательно, $z_\rho = 0{,}326 \pm 0{,}140$, откуда доверительный интервал для z_ρ имеет вид $(0{,}186\,;\ 0{,}466)$. Возвращаясь с помощью формулы (18.3) к переменной r, получим интервал $(0{,}184\,;\ 0{,}435)$, и, следовательно, с надёжностью $\gamma = 0{,}95$ можно утверждать, что коэффициент корреляции генеральной совокупности удовлетворяет неравенству $0{,}184 \le \rho \le 0{,}435$. Для удобства функция $z = \frac{1}{2}\cdot\ln\frac{1+r}{1-r}$ табулирована. Её значения приведены в таблице Н Приложения.

18.2. Проверка гипотезы $\rho_{xy} = a$. Проверяемая гипотеза состоит в том, что коэффициент корреляции Пирсона ρ между переменными X и Y равен некоторому числу a: $H_0\colon\ \rho_{xy} = a$ против $H_1\colon\ \rho_{xy} \ne a$. Предполагается, что генеральная совокупность, из которой *случайным* образом

выбираются пары $(x_i; y_i)$, является двумерной нормальной совокупностью, с коэффициентом корреляции ρ_{xy}. Для проверки H_0 против H_1:

а) уславливаются об уровне значимости α и объеме выборки n;

б) берут случайную выборку и вычисляют выборочный коэффициент корреляции r;

в) найденное значение r с помощью преобразования Фишера $\left(z_r = \frac{1}{2}\ln\frac{1+r}{1-r}\right)$ преобразуют в z_r. Как уже отмечалось, выборочное распределение для z_r почти нормально с дисперсией $1/(n-3)$;

г) для проверки нуль - гипотезы о том, что $\rho = a$, используется статистика:

$$z = \frac{z_r - z_a}{1/\sqrt{n-3}}, \qquad (18.4)$$

где z_r – преобразованное значение r; z_a – преобразованное значение a; n – объём выборки;

д) когда верна H_0, т.е. $\rho = a$, то выборочное распределение z в выражении (18.4) имеет стандартное нормальное распределение.

Пример. Многократные проверки пилотажного характера привели исследователя к предположению, что связь между оценками по алгебре и оценками по геометрии у восьмиклассников характеризуется коэффициентом корреляции Пирсона $\rho_{xy} = 0{,}70$.

Для того чтобы проверить гипотезу $H_0: \rho_{xy} = 0{,}70$ против $H_1: \rho_{xy} \neq 0{,}70$, исследователь случайным образом отобрал 25 человек и сопоставил их годовые оценки по алгебре и геометрии. Вычисленное выборочное значение r оказалось равным 0,78.

Для принятия или отклонения нулевой гипотезы необходимо, прежде всего, с помощью преобразования Фишера найти значения z_r и z_a:

$\hat{z}_r = z_{0,78} = \frac{1}{2}\ln\frac{1+0{,}78}{1-0{,}78} = 1{,}045$; $\hat{z}_a = z_{0,70} = \frac{1}{2}\ln\frac{1+0{,}70}{1-0{,}70} = 0{,}867$, после чего к этим значениям применить z-критерий (18.4): $\hat{z} = \frac{z_r - z_a}{1/\sqrt{n-3}} = \frac{1{,}045-0{,}867}{1/\sqrt{25-3}} = 0{,}83$.

Так как на уровне значимости $\alpha = 0{,}05$ критические точки z-критерия равны ${}_{0,975}z = -1{,}960$; ${}_{0,025}z = 1{,}960$, то найденное значение z=0,83 не по-

падает в критические области и, следовательно, на уровне значимости $\alpha = 0{,}05$ нет оснований отклонять нулевую гипотезу.

Доверительные интервалы для ρ_{xy} строятся путём определения доверительного интервала для z_ρ относительно z_r и последующего преобразования верхних и нижних пределов полученного интервала в исходную шкалу r с помощью преобразования, обратного фишеровскому.

В рассмотренном примере доверительный интервал для ρ_{xy} вычисляется в два этапа. Сначала находят доверительный интервал для z_ρ:

$$z_\rho = z_r \pm_{(1-\alpha/2)} z\frac{1}{\sqrt{n-3}} = 1{,}045 \pm 1{,}960\frac{1}{\sqrt{25-3}} = 1{,}045 \pm 0{,}418,$$

откуда $(0{,}627;1{,}463)$.

Затем, пользуясь обратным преобразованием Фишера: $r = \dfrac{e^{2z_r}-1}{e^{2z_r}+1}$, находят границы интервала, содержащего на уровне значимости $\alpha = 0{,}05$ значение ρ_{xy}: $(0{,}556\ ;\ 0{,}898)$.

Замечание. На практике в большинстве случаев проверяется гипотеза: $H_0 : \rho_{xy} = 0$. В этом случае критерий (18.4) приобретает вид $z = \dfrac{z_r}{1/\sqrt{n-3}}$**.** Более удобной в этом случае $(H_0 : \rho_{xy} = 0)$ является статистика

$$\hat{t} = \frac{r_{xy}}{\sqrt{\left(1-r_{xy}^2\right)/(n-2)}}, \qquad (18.5)$$

достаточно точно описываемая $t-$ распределением Стьюдента с числом степеней свободы $\nu = n-2$.

18.3. Проверка гипотезы $\rho_1 = \rho_2$ по независимым выборкам. Рассмотрим два статистических коллектива, например, группу мальчиков и группу девочек одного и того же возраста. Каждый из наблюдаемых характеризуется двумя показателями: способностью (X) и успеваемостью (Y). ρ_1 и ρ_2 – коэффициенты корреляции между X и Y в каждом из этих коллективов. Кроме того, предполагается, что каждый из коллективов имеет двумерное нормальное распределение относительно X и Y.

Нуль-гипотеза как обычно утверждает, что $\rho_1 = \rho_2$, альтернативная гипотеза – обратное: $H_0{:}\rho_1 = \rho_2$, $H_1{:}\rho_1 \neq \rho_2$. Для проверки нулевой гипотезы:

а) случайным образом выберем из первой группы n_1 мальчика и независимо от них из второй группы n_2 девочки;

б) для каждой выборки вычислим выборочный коэффициент корреляции r_1 и r_2;

в) найденные значения r_1 и r_2 с помощью z-преобразования Фишера преобразуем в z_{r_1} и z_{r_2};

г) располагая значениями z_{r_1} и z_{r_2}, обратимся к статистике:

$$\hat{z} = \frac{z_{r_1} - z_{r_2}}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}}, \qquad (18.6)$$

если $\rho_1 = \rho_2$, то z в выражении (18.6) имеет стандартное нормальное распределение.

Обращаясь к рассмотренному в начале этого пункта примеру, предположим, что в выборке из 94 девочек 15-ти лет связь между интеллектом (Стенфорд-Бине) X и средней успеваемостью по всем предметам Y определилась величиной $r_1 = 0{,}65$. В выборке из 80 мальчиков того же возраста связь между интеллектом и средней успеваемостью выразилась величиной $r_2 = 0{,}41$.

Нулевую гипотезу $H_0{:}\rho_1 = \rho_2$ проверим на уровне значимости $\alpha = 0{,}05$. Для этого, прежде всего преобразуем значения r_1 и r_2 в $z_{r_1} = 0{,}775$ и $z_{r_2} = 0{,}436$. Затем, применяя выражение (18.6), получим

$$\hat{z} = \frac{0{,}775 - 0{,}436}{\sqrt{1/91 + 1/77}} = 2{,}19.$$

Так как критические точки ${}_{0{,}025}z = -1{,}960$, ${}_{0{,}975}z = 1{,}960$, и z попадает в верхнюю критическую область, то нулевую гипотезу можно отклонить на уровне значимости $\alpha = 0{,}05$. 95%-ный доверительный интервал для $\rho_1 - \rho_2$ определяется следующим образом:

$z_{r_1} - z_{r_2} = 0{,}775\text{-}0{,}436 \pm 1{,}96 \cdot 2{,}19 = 0{,}339 \pm 0{,}303$ или (0,036; 0,642).

Преобразование этих двух z-величин снова в r-шкалу дает 95%-ный доверительный интервал для $\rho_1 - \rho_2$: (0,036; 0,566).

18.4. Оценка линейности регрессии по корреляционной таблице. Если данные представлены в виде корреляционной таблицы, то обычно используется корреляционное отношение E^2_{yx} (см. пункт 17.3), которое характеризует степень отклонения частоты по столбцам от средних значений по столбцам:

$$r^2 \le E^2_{yx} \le 1. \qquad (18.7)$$

При линейной регрессии корреляционное отношение и коэффициент корреляции примерно равны друг другу. Чем больше отклонение средних значений по столбцам от прямой, тем больше разность между E_{yx} и r. Эту разность между статистиками можно использовать для проверки гипотезы линейности.

Отношение

$$F = \frac{\frac{1}{k-2}(E^2_{yx} - r^2)}{\frac{1}{n-k}\left(1 - E^2_{yx}\right)}; \qquad \begin{matrix} \nu_1 = k-2, \\ \nu_2 = n-k, \end{matrix} \qquad (18.8)$$

где $k-$ число столбцов, подчиняется F-распределению с $\nu_1 = k-2$ и $\nu_2 = n-k$ степенями свободы. Значимое F-отношение соответствует значимому отклонению от линейности.

В качестве примера обратимся к корреляционной таблице 17,3. Проведенные в пункте 17.3 расчеты показали, что E^2_{yx}=0,70, а $r^2_{xy} = -0{,}035$. Обращаясь к статистике (18.9) и учитывая, что k=8, а n=37, получим

$$\widehat{F} = \frac{\frac{1}{8-2}\left(0{,}70 + 0{,}035^2\right)}{\frac{1}{37-8}\left(1 - 0{,}70\right)} = \frac{0{,}11667}{0.010345} = 11{,}28.$$

Так как $\widehat{F} = 11{,}28 > 3{,}5 = {}_{0{,}01}F_{6,29}$ (табл. D_2 Приложения), есть все основания отвергнуть нулевую гипотезу.

18.5. Проверка значимости коэффициента регрессии. Если проверка, проведенная по указанному выше правилу, не дает основания сомневаться в линейности регрессии и ее уравнение имеет вид $\widetilde{y} = ax + b$, где $a = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$, $b = \bar{y} - a\bar{x}$ (пункт 17.1), то может быть осуществлена проверка значимости коэффициента регрессии, а именно проверяется нуль-гипотеза H_0: $a = 0$, т.е. проверяется, отличается ли статистически

значимо оценка коэффициента регрессии (a) от нуля. Граница значимости устанавливается на основании распределения Стьюдента

$$\hat{t} = \frac{|a|}{S_a}\ , \quad \text{с (n-2) степенями свободы,} \tag{18.10}$$

где $S_a = \dfrac{\sqrt{\overline{y^2} - \bar{y}^2 - a\left(\overline{xy} - \bar{x}\cdot\bar{y}\right)}}{\sqrt{n-2}\sqrt{\overline{x^2} - \bar{x}^2}}$.

Если статистика больше, чем граница значимости или равна ей, то a значимо отличается от нуля.

В качестве примера рассмотрим корреляционную зависимость, определяемую таблицей 18.1

Пример.

*Таблица 18.1.*Корреляционная таблица

a_y	Значения a_x						n_y	$n_y \cdot a_y$	$n_y \cdot a_y^2$
	-3	-2	-1	0	1	2			
1			1	5	7	1	14	14	14
0		1	3	7	5	2	18	0	0
-1		2	3	4	1		10	-10	10
-2		3	1	1			5	-10	20
-3	2	1					3	-9	27
$n_{i\cdot}$	2	7	8	17	13	3	50	-15	71
$n_{i\cdot}x_i$	-6	-14	-8	0	13	6	-9		
$n_i x_i^2$	18	28	8	0	13	12	70		

Из приведенных в таблице данных следует, что $\bar{x}$=-9/50=- 0,18, $\bar{y}$=-15/50=- 0,30, $\overline{x^2}$=70/50 =1,40, $\overline{y^2}$=71/50 =1,42, $\overline{xy}$=52/50=1,04.

$$\hat{r} = \frac{1{,}04 - 0{,}18\cdot 0{,}3}{\sqrt{1{,}4 - 0{,}18^2}\sqrt{1{,}42 - 0.3^2}} = \frac{0{,}986}{1{,}1694\cdot 1{,}1533} = 0{,}731,$$

$$E_{yx}^2 = \frac{S_1 - R}{S_2 - R}, \text{ где } S_1 = \frac{(-6)^2}{2} + \frac{(-1)^2}{7} + \frac{(-4)^2}{8} + \frac{(-1)^2}{17} + \frac{6^2}{13} + \frac{1^2}{3} = 40{,}45,$$

$S_2 = 71,\ R = (-15)^2/50 = 4{,}5$, откуда $E_{yx}^2 = \dfrac{40{,}45 - 4{,}5}{71 - 4{,}5} = 0{,}541$,

$$\hat{F} = \frac{\dfrac{1}{6-2}\left(0{,}541 - 0{,}731^2\right)}{\dfrac{1}{50-6}\left(1 - 0{,}541\right)} = 1{,}653.$$

Так как $\widehat{F} = 1{,}653 < 2{,}55 = {}_{0{,}05}F_{4,54}$, то нет оснований отвергать гипотезу о линейности.

$$\widehat{a} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{1{,}04 = 0{,}18 \cdot 0{,}30}{1{,}40 - 0{,}18^2} = \frac{0{,}986}{1{,}3676} = 0{,}721;$$

$\widehat{b} = \bar{y} - a\bar{x}$ =0,30–0,721·0,18=0,17. Линия регрессии y=0,72x+0,17.

В заключение оценим значимость коэффициента регрессии. С этой целью вычислим S_a. По формуле (18.10) получим

$$\widehat{S}_a = \frac{\sqrt{\overline{y^2} - \bar{y}^2 - a\left(\overline{xy} - \bar{x} \cdot \bar{y}\right)}}{\sqrt{n-2}\sqrt{\overline{x^2} - \bar{x}^2}} \frac{\sqrt{1{,}42 - 0{,}09 - 0{,}721\left(1{,}04 - 0{,}18 \cdot 0{,}3\right)}}{\sqrt{50-2}\sqrt{1{,}4 - 0{,}18^2}} = 0{,}097$$

$\widehat{t} = \frac{|a|}{s_a} = \frac{0{,}721}{0{,}097} = 7{,}43 > 2.011 = {}_{0{,}05}t_{50}$, следовательно a значимо отличается от нуля.

Глава 19. Частные виды коэффициента корреляции

19.1. Классификация коэффициентов корреляционных связей

Применение и интерпретация корреляционного отношения E^2_{yx} и коэффициента корреляции Пирсона r_{xy} зависят от характера данных, которые находятся в корреляционной связи: будут ли они соответствовать двумерному нормальному распределению, могут ли как X, так и y принимать непрерывные значения, будут ли два распределения иметь идентичную форму.

Теперь следует приступить к исследованию мер связи, применяемых к переменным, не столь хорошо оцениваемым количественно, как вес, возраст, время выполнения задания и др.

В частности, будут рассмотрены коэффициенты, оценивающие степень корреляционной связи между переменными, измеренными в дихотомических или порядковых шкалах. Некоторые коэффициенты появятся в результате применения формулы $r_{xy}=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2}-\bar{x}^2}\cdot\sqrt{\overline{y^2}-\bar{y}^2}}$ непосредственно к новым видам данных. Другие будут введены при попытке оценить, каким бы было значение r_{xy}, если бы данные не были представлены в необработанном виде. Кроме того, в этой главе будут рассмотрены коэффициенты, основанные на иных, чем прежде, представлениях о том, что такое «связь» и как её следует измерять.

Характер коэффициента корреляционной связи во многом зависит от типа измерения переменных X и Y.

До сих пор мы рассматривали коэффициенты связи между переменными, измеренными в шкалах интервалов или отношений. Теперь мы предположим, что одна или обе переменных измерены в шкалах более низкого уровня: дихотомической или порядковой. Кроме того, в ряде случаев будем предполагать, что более полные и более совершенные методы измерения могли бы обеспечить приблизительно нормальное распределение результатов. Сочетание четырёх типов шкал (табл. 19.1) создаёт в общей сложности 16 ситуаций, из которых принципиально различными будут десять. Объем пособия не позволяет рассмотреть их все. Мы ограничимся коэффициентами, наиболее часто применяемыми в психолого-педагогических и социологических исследованиях.

Таблица 19.1. Классификация коэффициентов корреляционных связей

	Шкала переменной X			
Шкала переменной Y	дихотомическая шкала	дихотомия, основанная на нормальном распределении	шкала порядка	шкала интервалов или отношений
Дихотомическая шкала	A	B	C	D
Дихотомия, основанная на нормальном распределении	B	E	F	H
Шкала порядка	C	F	G	L
Шкала интервалов или отношений	D	H	L	K

В частности, случай A представлен коэффициентом φ (пункт 19.2), случай E – тетрахорическим коэффициентом корреляции r_{tet} (пункт 19.3). Случай *D* – точечно-бисериальным коэффициентом корреляции r_{pb} (пункт 19.4). Случай *H* – бисериальным коэффициентом корреляции r_{bis} (пункт 19.5).

Наконец, случай G представлен коэффициентом ранговой корреляции Спирмена r_s (пункт 20.1). Хотя этот коэффициент формально выводится из коэффициента Пирсона, он обладает рядом качеств, которые делают его типично непараметрическим критерием, малочувствительным к типу распределения и системе мер. Применение коэффициента Спирмена дает достаточно точный результат в случае малого объема выборки и при ее нормальном законе распределения; кроме того, ослабляется влияние выбросов, которые могут сильно изменять значение коэффициента корреляции.

Рассмотрение некоторых важных критериев оценки корреляции завершается в двадцатой главе (пункты 20.3 и 20.4) рассмотрением развитой на основании углового критерия медианной или квадрантной корреляции, пригодной для быстрой ориентации. Подобно коэффициенту ранговой корреляции, коэффициент квадрантной корреляции позволяет проводить надежную проверку при любой функции распределения, уменьшить влияние выбросов, он независим от системы мер.

Большое число различных коэффициентов корреляции соответствует большому разнообразию типов связей в природе и обществе. Оно обеспечивает необходимую приспособляемость статистического аппарата к анализу сложнейших взаимосвязей в социальной области. Каждый корре-

ляционный коэффициент приспособлен для измерения вполне определённого вида связи. Применение того или иного показателя определяется природой данных и формой их представления. Учитывается также и требуемая степень точности.

Обычно стараются использовать наиболее распространённые в практике психолого-педагогических и социологических исследований коэффициенты, так как это обеспечивает возможность сравнивать полученные результаты с материалами других исследователей.

19.2. Коэффициент φ для собственно дихотомических шкал. Рассмотрим ситуацию, обозначенную в таблице 19.1 буквой А, т.е. случай, когда оба признака «измерены» в номинальной шкале с двумя значениями, обозначенными нулем и единицей. Начнем с примера.

Пример. Изучается связь между полом студентов и пристрастием к курению. С этой целью из списка студентов факультета случайным образом извлекается выборка объёмом $n = 100$ человек, и отмечаются те, кто считает себя курящим. Результаты опроса оформляются корреляционным рядом, в котором X означает пол респондента (0 – женщина, 1– мужчина), а Y – факт курения (0 – не курит, 1 – курит). Начало этого ряда приведено в таблице 19.2.

Таблица 19.2. Фрагмент данных, характеризующих связь между пристрастием к курению и полом студента

Студент	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	R	S
Пол, X	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0
Курение, Y	0	1	1	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0

Неудобство такого представления корреляционного ряда и его статистической обработки при большом объеме выборки очевидно. Поэтому полученные при опросе данные оформляют в виде таблицы (19.3).

Таблица 19.3. Связь между полом X и курением Y (в людях)

Курение, Y		Пол, X		Итого
		мужской (1)	женский (0)	
Курят	1	15	18	33
Не курят	0	5	62	67
Итого		20	80	100

Обращаясь к общему случаю, корреляционный ряд дихотомических переменных всегда можно представить в виде таблицы сопряженности 19.4.

Таблица 19.4. Таблица сопряженности признаков X и Y

		Признак X		Итого
		1	0	
Признак Y	1	a	b	$a+b$
	0	c	d	$c+d$
Итого		$a+c$	$b+d$	n

Используя коэффициент корреляции Пирсона, который в этом случае обозначают буквой φ, получим:

$$\varphi = r_{xy} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\cdot\sqrt{\overline{y^2} - \bar{y}^2}} = \frac{\frac{a}{n} - \frac{a+c}{n}\cdot\frac{a+b}{n}}{\sqrt{\frac{a+c}{n} - \left(\frac{a+c}{n}\right)^2}\sqrt{\frac{a+b}{n} - \left(\frac{a+b}{n}\right)^2}} =$$

$$= \frac{na - (a+c)(a+b)}{\sqrt{n(a+c) - (a+c)^2}\sqrt{n(a+b) - (a+b)^2}}.$$ Заменив n на сумму $(a+b+c+d)$, раскрыв скобки и приведя подобные члены, получим:

$$\varphi = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}. \qquad (19.1)$$

Так как знак правой части равенства зависит от случайных обстоятельств – от того, каким значениям переменных X и Y мы приписываем нуль, а каким – единицу, поэтому, как правило, оперируют значением $|\varphi|$.

Таким образом, коэффициент φ представляет собой просто коэффициент Пирсона для дихотомических данных. Однако, как мы увидим ниже, интерпретация φ может выдвигать специфические проблемы.

Используя выведенную формулу (19.1), вычислим коэффициент корреляции между полом студента и его пристрастием к курению на рассматриваемом факультете (табл. 19.3): $\hat{\varphi} = \frac{15\cdot 62 - 18\cdot 5}{\sqrt{20\cdot 80\cdot 33\cdot 67}} = 0{,}447$.

Обратимся теперь к статистической оценке коэффициента φ. Рассматривается генеральная совокупность, каждый член которой характеризуется двумя дихотомическими переменными X и Y. Если коэффициент связи φ этих переменных в генеральной совокупности равен нулю, то для достаточно больших выборок (n>20) выборочное распределение статистики

$$z = \sqrt{n}\cdot\varphi \qquad (19.2)$$

приближённо описывается стандартизированным нормальным законом.

Замечание. Когда коэффициент фи генеральной совокупности отличается от нуля, выборочное распределение статистики $\sqrt{n}\cdot\varphi$ становится скошенным, группируясь около среднего, отклоняющегося от нулевого значения на величину, которая увеличивается с удалением значений φ от нуля.

В рассмотренном примере $\hat{z}=\sqrt{n}\cdot\varphi=\sqrt{100}\cdot 0{,}447=4{,}47$.

Условившись об уровне значимости $\alpha=0{,}05$ и имея в виду, что критерий z имеет стандартизированное нормальное распределение, найдём критические значения ${}_{0{,}975}z=-1{,}960$ и ${}_{0{,}025}z=1{,}960$. Так как выборочное значение $z=4{,}47$ больше верхнего критического значения ${}_{0{,}025}z=1{,}960$, то гипотезу, будто бы «пол» и «курение» не связаны в рассматриваемой совокупности (студентов данного факультета), можно отклонить на уровне значимости $\alpha=0{,}05$.

19.3. Тетрахорический коэффициент корреляции. Рассмотрим теперь случай, когда две переменные X и Y, измеренные, как и в предыдущем случае, в дихотомических шкалах (0 и 1), на самом деле являются непрерывными величинами, которые при более дорогостоящих и широких действиях могли бы быть измерены в интервальных шкалах или шкалах отношений (случай Е таблицы 19.1). Так, например, вместо точного значения роста каждому испытуемому можно приписывать нуль, если его рост меньше среднего и единицу, если его рост больше среднего.

Результаты таких измерений могут быть оформлены таблицей, аналогичной таблице 19.3, и для них может быть вычислен коэффициент φ. Однако исследователь в этом случае может рассчитывать на более высокое значение связи между X и Y по сравнению с тем, которое фиксируется коэффициентом φ.

В этом случае может быть использован так называемый тетрахорический коэффициент корреляции r_{tet} между свойствами X и Y. Однако, точная формула для r_{tet} очень сложна, поэтому на практике используют более удобную, хотя и менее точную аппроксимацию:

$$r_{tet}=\cos\frac{180^0}{1+\sqrt{bc/ad}}. \qquad (19.3)$$

Заметим, что, как и в случае вычисления φ, изменение обозначений в одной из переменных нулей на единицы, и наоборот, которое ведёт к из-

менению отношения bc/ad на отношение ad/bc, приводит к изменению знака r_{tet}, т.е. $\cos\frac{180^\circ}{1+\sqrt{ad/bc}} = -\cos\frac{180^\circ}{1+\sqrt{bc/ad}}$.

Замечание. Не следует применять эту формулу, если отношения $(a+b)/n$ или $(a+c)/n$ лежат вне интервала $(0{,}30\,;\,0{,}70)$, так как это может привести к существенным ошибкам.

Пределами величины r_{tet} в силу строения формулы 19.3 служат числа -1 и 1 независимо от того, насколько неравны между собой числа $(a+b)/n$ и $(b+d)/n$. Это свойство заставляет предпочесть r_{tet} коэффициенту φ в качестве меры связи в случае, когда можно предположить, что в основе дихотомии лежат нормальные распределения.

Пример. Предположим, что изучается связь между ростом юношей 17 лет (X) и их весом (Y). При этом для «измерения» роста устанавливают мерную планку на высоте 170 см и приписывают нуль тем юношам, которые оказываются ниже планки, и единицу тем, чей рост оказывается выше. Взвешивание производят аналогично: устанавливают на весах 60 кг и отмечают нулём тех юношей, которые не дотягивают до 60 кг и единицей тех, чей вес оказывается выше или равен 60 кг. Фрагмент результатов измерения приведён в таблице 19.5.

Таблица 19.5. Зависимость между ростом и весом юношей 17 лет

Y - вес (в кг)		X – рост (в см)		Итого
		больше или равен 170	меньше 170	
		1	0	
Больше или равен 60	1	a = 20	b = 10	30
Меньше 60	0	c = 5	d = 25	30
Итого		25	35	60

Так как рост и вес являются непрерывными величинами, которые при более дорогостоящих и широких действиях могли бы быть измерены в шкалах отношений, а значения $(a+b)/n = 0{,}5$, и $(a+c)/n = 0{,}42$ не выходят за пределы интервала $(0{,}30;\ 0{,}70)$, то представляется возможным использование тетрахорического коэффициента для определения связи между ростом и весом юношей 17 лет:

$$\hat{r}_{tet} = \cos\frac{180^\circ}{1+\sqrt{bc/ad}} = \cos\frac{180^\circ}{1+\sqrt{500/50}} = \frac{180^\circ}{1+\sqrt{10}} = 0{,}73.$$

Если бы мы воспользовались коэффициентом φ, то получили бы значение $\hat{\varphi} = \dfrac{500 - 50}{\sqrt{25 \cdot 35 \cdot 30 \cdot 30}} = 0{,}51$, которое существенно меньше r_{tet}.

Для умеренно больших и больших n для проверки гипотезы о том, что r_{tet} генеральной совокупности равен нулю, применяется статистика

$$z = r_{tet} / \sigma_{tet}, \qquad (19.4)$$

где $\sigma_{tet} = \sqrt{\dfrac{p_x p_y q_x q_y}{n}} \cdot \dfrac{1}{u_x u_y}$.

Здесь n – объём выборки; $p_x = (a+c)/n$ – доля объектов, имеющих единицу по дихотомической переменной X; $p_y = (a+b)/n$ – доля объектов с единицами по дихотомической переменной Y; $q_x = 1 - p_x$, $q_y = 1 - p_y$, наконец, u_x – ордината стандартной нормальной кривой в точке x, правее которой находится доля p_x всех ее значений; u_y - ордината стандартной нормальной кривой в точке x, правее которой находится доля p_y всех ее значений.

Если верна нулевая гипотеза о равенстве нулю тетрахорического коэффициента корреляции генеральной совокупности, то статистику $z = r_{tet} / \sigma_{r_{tet}}$ можно отнести к стандартному нормальному распределению.

Для расчёта $\sigma_{r_{tet}}$ найдём $p_x = \dfrac{25}{60} = 0{,}417$; $q_x = 1 - 0{,}417 = 0{,}583$; $p_y = \dfrac{30}{60} = 0{,}50$; $q_y = 0{,}50$. Для нахождения значений u_x и u_y обратимся к таблице значений стандартной нормальной кривой (таблица N Приложения). По этой таблице ордината стандартной нормальной кривой в точке $(x = 0{,}21)$, правее которой находится доля $p_x = 0{,}417$ всех ее значений, равна $u_x = 0{,}3902$; ордината стандартной нормальной кривой в точке $(x = 0)$, правее которой лежит доля $p_y = 0{,}50$ всех ее значений, равна $u_y = 0{,}3989$. Откуда значение z критерия (19.4) равно:

$$\hat{z} = \frac{0{,}73}{\sqrt{\dfrac{0{,}417 \cdot 0{,}50 \cdot 0{,}583 \cdot 0{,}50}{60}} \cdot \dfrac{1}{0{,}3902 \cdot 0{,}3989}} = 3{,}57.$$

Таким образом, нулевая гипотеза о равенстве нулю тетрахорического коэффициента генеральной совокупности может быть отклонена даже на уровне значимости $\alpha = 0{,}01$, т.к. $t = 3{,}57$ больше ${}_{0{,}01}t_{58} = 2{,}660$.

19.4. Точечно-бисериальный коэффициент корреляции. Рассмотрим теперь ситуацию, обозначенную в таблице 19.1 буквой D, т.е. когда одна из переменных измерена в шкале интервалов или отношений, а другая в дихотомической шкале.

Как и в предыдущем случае, начнём с примера. Рассмотрим фрагмент таблицы, выражающей зависимость между ростом школьника и его полом (табл. 19.6).

Таблица 19.6. Связь между ростом и полом десятиклассников. X – рост в см., Y – пол (1– юноша, 0 – девушка).

Ученики	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Итого
X – рост	170	150	160	165	140	183	157	152	163	168	170	145	167	150	152	148	2540
Y – пол	1	0	1	1	0	1	0	0	1	1	1	0	1	0	0	0	8

Для описания связи X и Y можно использовать коэффициент корреляции Пирсона в предположении, что значения Y, как и значения X, измерены в шкале отношений. Вычисленный при таком предположении коэффициент корреляционной связи называют точечным бисериальным коэффициентом корреляции (бисериальным произведением моментов) и обозначают r_{pb}.

В нашем примере:

$$\bar{x} = 158{,}75,\ \bar{y} = 0{,}5,\ \overline{xy} = 84{,}126,\ \overline{x^2} = 25323{,}88,\quad \overline{y^2} = 0{,}5,$$

$$\hat{r}_{pb} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\cdot\sqrt{\overline{y^2} - \bar{y}^2}} = \frac{84{,}125 - 158{,}75\cdot 0{,}5}{\sqrt{25324 - 158{,}75^2}\cdot\sqrt{0{,}5 - 0{,}5^2}} = \frac{4{,}75}{5{,}53} = 0{,}86.$$

Вычисления могут быть несколько упрощены, если в формулу Пирсона внести изменения, обусловленные особенностью значений переменной Y.

Введём обозначения:

n_1 – число объектов наблюдения, имеющих по признаку Y единицу;

n_0 – число объектов наблюдения, имеющих по признаку Y нуль;

$\bar{n}_1$ – среднее по объектам из X, имеющих по признаку Y единицу;

$\bar{n}_0$ – среднее по объектам из X, имеющих по признаку Y нуль.;

S_x – стандартное отклонение всех n значений из X.

Тогда $n_0 + n_1 = n$ – общее число объектов наблюдения,

$$\bar{x} = \frac{n_0\bar{x}_0 + n_1\bar{x}_1}{n}, \quad \bar{y} = \frac{n_1}{n}, \quad \overline{y^2} = \frac{n_1}{n}, \quad \overline{xy} = \frac{\bar{x}_1 n_1}{n} \quad \text{и} \quad \overline{x^2} - \bar{x}^2 = \frac{n-1}{n} \cdot S_x^2$$

(см. формулу 5.3). Подставив эти значения в формулу коэффициента корреляции Пирсона $r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}$ и выполнив соответствующие преобразования, получим выражение для точечного бисериального коэффициента корреляции

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{S_x} \cdot \sqrt{\frac{n_0 \cdot n_1}{n(n-1)}}. \tag{19.5}$$

В рассмотренном выше примере (таблица 19.5) n=16, $n_1 = 8$, n_0=8, $\bar{x}_1 = 168{,}25$, $\bar{x}_0 = 149{,}25$, $\bar{x} = 158{,}75$, $\overline{x^2} = 25323{,}88$, откуда $S_x = 11{,}422$.

Поэтому $\hat{r}_{pb} = \frac{168{,}25 - 149{,}25}{11{,}422} \cdot \sqrt{\frac{8 \cdot 8}{16 \cdot 15}} = 1{,}663 \cdot 0{,}516 = 0{,}86$.

Проверка нулевой гипотезы $H_0: r_{pb} = 0$ против $H_1: r_{pb} \neq 0$ осуществляется с помощью статистики

$$t = \frac{r_{pb}}{\sqrt{\left(1 - r_{pb}^2\right)/(n-2)}}, \tag{19.6}$$

которая приближённо соответствует t-распределению Стьюдента с $n-2$ степенями свободы.

Применяя этот признак к нашему случаю, получим $t = \frac{0{,}86}{\sqrt{\left(1 - 0{,}86^2\right)(16-2)}} = \frac{0{,}86}{0{,}1364} = 6{,}30$. Так как ${}_{0{,}05}t_{14} = 2{,}140$, то результаты проведенного эксперимента дают основание для отклонения нульгипотезы об отсутствии значимой корреляционной связи между полом школьника и его ростом.

19.5. Бисериальный коэффициент корреляции. Рассмотрим ситуацию, отмеченную в таблице 19.1 буквой Н, когда одна из переменных измеряется дихотомически, на основе нормального распределения, а другая – в шкале интервалов или отношений. В качестве примера рассмотрим связь между временем X, затрачиваемым на изучение материала, и уровнем Y его усвоения, измеряемом в дихотомической шкале по ответам на

контролирующие вопросы (1 – если ответ верен, 0 – если ответ неверен). Фрагмент полученного в результате эксперимента корреляционного ряда приведён в таблице 19.7.

Таблица 19.7. Связь между временем (в минутах), затрачиваемом на изучение материала, и уровнем его усвоения: 1 – если ответ верен и 0 – если неверен

Ученики	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	R	S
X – время	16	12	11	7	15	14	10	11	15	9	13	7	13	11	10	11	10	11
Y – ответ	1	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	0	1

Мерой связи служит так называемый бисериальный коэффициент корреляции r_{bis}, в некотором смысле аналогичный коэффициенту r_{tet}, за исключением того, что для r_{tet} обе переменные были дихотомическими.

Формула для вычисления r_{bis} выводится на основе теории регрессии и имеет вид:

$$r_{bis} = \frac{\bar{x}_1 - \bar{x}_0}{S_x} \cdot \frac{n_1 n_0}{u \cdot n\sqrt{n^2 - n}}. \qquad (19.7)$$

В ней $\bar{x}_1$ и $\bar{x}_0$ являются средними арифметическими значениями X для объектов, имеющих по Y соответственно 1 и 0; S_x – стандартное отклонение значений X; n_1 и n_0 – число единиц и нулей в Y; $n = n_1 + n_0$: u – ордината (т.е. высота) нормированного нормального распределения в точке, за которой лежит $\frac{n_1}{n}100$ процентов площади под кривой.

В рассмотренном выше примере $n_1 = 10$; $n_0 = 8$; $n = 18$; $\bar{x}_1 = 12,6$; $\bar{x}_0 = 10,0$; $\bar{x} = \frac{10 \cdot 12,6 + 8 \cdot 10}{18} = 11,444$, $\overline{x^2} = 137,111$, $S_x = \sqrt{\frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right)}$ $\sqrt{\frac{18}{17} \cdot \left(137,11 - 11,44^2\right)} = 2,5489.$

Для нахождения параметра u учтём, что $\frac{n_1}{n} = \frac{10}{18} = 0,556$, и, пользуясь таблицей Н Приложения, найдём ординату u нормированной нормальной кривой для точки, правее которой лежит 55,6 процентов всей площади. Для этого по таблице значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x e^{-\frac{1}{2}t^2}\, dt$ (таблица N Приложения) находим $x = -0,14$, при котором

$\Phi(x) = 0{,}556 - 0{,}500 = 0{,}056$. В этой точке $\varphi(x) = 0{,}3951$. Таким образом, $u = 0{,}3951$.

Теперь можем вычислить

$$\hat{r}_{bis} = \frac{\bar{x}_1 - \bar{x}_0}{S_x} \cdot \frac{n_1 n_0}{u \cdot n \cdot \sqrt{n^2 - n}} = \frac{12{,}6 - 10{,}0}{2{,}54885} \cdot \frac{10{,}8}{0{,}3951 \cdot 18\sqrt{18^2 - 18}} = 0{,}656.$$

В отличие от большинства коэффициентов корреляции r_{bis} иногда может принимать значения ниже -1 и выше +1. Но это лишь означает, что либо некорректно предположение о нормальности распределения X, либо имеется флуктуация выборки, когда n мало ($n \le 15$).

Проверка нулевой гипотезы $H_0 : r_{bis} = 0$ против $H_1 : r_{bis} \neq 0$ осуществляется с помощью статистики

$$z = r_{bis} / \sigma_{r_{bis}}, \qquad (19{,}8)$$

где $\sigma_{r_{bis}} = \frac{\sqrt{n_0 \cdot n_1}}{u \cdot n \cdot \sqrt{n}}$.

Используя найденные выше значения, получим, что

$\hat{\sigma}_{r_{bis}} = \frac{\sqrt{10 \cdot 8}}{0{,}3951 \cdot 18 \cdot \sqrt{18}} = 0{,}296$, откуда $\hat{z} = \frac{0{,}656}{0{,}296} = 2{,}216$.

Сравнивая это значение с критическими точками стандартизированного нормального распределения, заключаем, что результаты эксперимента позволяют на уровне статистической значимости $\alpha = 0{,}05$ отклонить нулевую гипотезу об отсутствии связи между временем, затрачиваемом на изучение материала и уровнем его усвоения.

Глава 20. Непараметрические меры зависимости

20.1. Общие замечания. При выводе коэффициента корреляции Пирсона предполагалось, что рассматриваемая двумерная выборка извлекается из *бинормальной генеральной совокупности* с параметром ρ. В условиях реального эксперимента это требование нередко частично или полностью не выполняется. В таких случаях используется обычно без каких-либо преобразований и со значительной экономией времени *коэффициент ранговой корреляции Спирмена* (r_s) проверка при этом дает достаточно точный результат в случае малого объема выборки и при ее нормальном законе распределения; кроме того, ослабляется влияние выбросов, которые могут сильно изменить значение коэффициента корреляции (r).

Другим преимуществом рангового коэффициента корреляции является его *независимость от системы мер*, так как он, в противоположность обычному коэффициенту корреляции, не изменяет своего значения, когда при неизменной последовательности вместо значений x применяется монотонная функция $F(x)$.

Для больших выборок из бинормального распределения с достаточно малым коэффициентом корреляции ($|\rho|<0{,}25$) применение коэффициента r_s приводит к тем же результатам, что и применение коэффициента Пирсона r в выборке, содержащей 91% наблюдений. Ввиду небольших потерь в точности, при значительной экономии времени коэффициент ранговой корреляции Спирмена r_S используется для быстрой предварительной оценки обычного коэффициента корреляции. Если имеется нормальное распределение, то оценка значения $|\rho|$ несколько завышается. Хотя с увеличением объема выборки коэффициент r_S стремится не к ρ (как r), а к ρ_s, разница между ρ и ρ_S не превышает 1,8% от значения ρ.

Значительные преимущества имеет применение коэффициента r_S при нелинейной монотонной регрессии: например, когда между признаками имеется логарифмическая или экспоненциальная зависимость и когда при увеличении одной переменной другая в среднем или непрерывно возрастает либо непрерывно падает. Применение коэффициента r в качестве меры корреляции требует преобразования переменных, при котором взаимозависимость становится линейной, поэтому использование коэффициента r_S здесь приводит к значительной экономии времени.

Очень удобна также развитая на основании углового критерия медианная, или *квадрантная*, *корреляция по Кеную*, пригодная для быстрой ориентации. При нормальном распределении можно коэффициент квадрантной корреляции (r_Q) принять для оценки обычного коэффициента корреляции ρ. Правда, критерий r_Q в этом случае недостаточно строгий, так как он охватывает только 41% всех наблюдений.

Подобно коэффициенту ранговой корреляции, коэффициент квадрантной корреляции имеет преимущества: он позволяет проводить надежную *проверку* при любой функции распределения, *уменьшать влияние выбросов* и является *независимым от системы мер*.

Обратимся теперь к описанию конкретных непараметрических мер зависимости.

20.2. Коэффициент ранговой корреляции Спирмена. Обратимся теперь к ситуации (G), когда обе переменные измерены в шкалах порядка. Исходные данные могут быть преобразованы в ранги или изначально быть рангами, полученными в результате экспертного оценивания.

В этом случае для построения коэффициента, характеризующего степень корреляционной связи, используют коэффициент корреляции Пирсона, вычисляемый по двум группам n последовательных, *несвязанных* рангов $1, 2, \ldots, n$.

В качестве примера рассмотрим таблицу 20.1, в которой приведены результаты ранжирования учащихся по двум признакам: успеваемости по математике и физике.

Таблица 20.1. Ранжирование учащихся по двум признакам: успеваемости по математике (X) и успеваемости по физике (Y).

	A	B	C	D	E	F	G	H	I	J	K	L
Успеваемость по математике (X)	4	8	12	5	1	6	7	10	2	9	11	3
Успеваемость по физике (Y)	6	5	10	7	3	4	9	8	1	11	12	2

Рассматривая каждый из рангов как числовое значение величин X и Y в шкале отношений, вычислим коэффициент корреляции Пирсона. Легко проверить, что $\bar{x} = \frac{78}{12} = 6{,}5$; $\bar{y} = \frac{78}{12} = 6{,}5$. Это совпадение не случайно. Ведь в конечном счете ищется среднее арифметическое чисел $1, 2, \cdots, n$. Поэтому при n рангах $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = 1 + 2 + \cdots + n = \frac{1+n}{2} \cdot n$, а $\bar{x} = \bar{y} = \frac{n+1}{2}$.

Аналогично для $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$,

откуда $\overline{x^2} = \overline{y^2} = \frac{(n+1)(2n+1)}{6}$.

Для нахождения $\overline{xy}$ рассмотрим равенство $\sum_{i=1}^{n} (x_i - y_i)^2 =$

$= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2$, откуда $\sum_{i=1}^{n} x_i y_i = \frac{1}{2}\left[\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} (x_i - y_i)^2\right]$.

Разделив обе части последнего равенства на n, получим, что

$$\overline{xy} = \frac{\overline{x^2} + \overline{y^2}}{2} - \frac{\sum (x_i - y_i)^2}{2n} = \frac{(n+1)(2n+1)}{6} - \frac{\sum (x_i - y_i)^2}{2n}.$$

Подставив найденные значения $\bar{x}, \bar{y}, \overline{x^2}, \overline{y^2}, \overline{xy}$ в формулу $r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}$, после очевидных преобразований получим:

$$r_{xy} = 1 - \frac{6 \cdot \sum (x_i - y_i)^2}{n(n^2 - 1)}.$$

Полученный коэффициент называют коэффициентом ранговой корреляции по Спирмену и обозначают r_s. Обозначив разность $x_i - y_i = d_i$, получим

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}. \qquad (20.1)$$

Таким образом, для вычисления коэффициента ранговой корреляции по Спирмену достаточно вычислить $\sum d_i^2$. В нашем случае такой расчёт выполнен в таблице 20.2.

Таблица 20.2. Расчет рангового коэффициента корреляции

x_i	4	8	12	5	1	6	7	10	2	9	11	3
y_i	6	5	10	7	3	4	9	8	1	11	12	2
d_i^2	4	9	4	4	4	4	4	4	1	4	1	1

$\sum d_i^2 = 44$, $\hat{r}_s = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 44}{12(144 - 1)} = 0{,}85$.

Коэффициент r_S нельзя истолковать иначе, чем сказать, что он равен коэффициенту корреляции Пирсона, вычисленному по рангам. Его значе-

ния, как и значения r_{xy}, могут меняться в пределах от -1 до 1. Коэффициент ранговой корреляции Спирмена особенно удобен тогда, когда исходные данные представляют собой ранги, например, когда эксперты ранжируют людей, события или предметы. Он иногда рассматривается как средство быстрой оценки r_{xy}. В этом случае при ранжировании численных значений переменной может возникнуть ситуация, когда ранги надо приписать двум одинаковым значениям переменной. В подобных случаях обоим значениям присваивается ранг, равный среднему арифметическому двух очередных рангов.

Выборочное распределение r_S, характеризующее связь между двумя группами рангов X и Y при $n \le 10$ нельзя описать в терминах любых хорошо известных распределений.

Если коэффициент ранговой корреляции Спирмена генеральной совокупности равен нулю, то выборочное распределение r_S, при $n > 10$, может быть связано с t - распределением Стьюдента следующей формулой:

$$t = \frac{r_s}{\sqrt{\left(1 - r_n^2\right)/(n-2)}} \qquad (20.2)$$

с числом степеней свободы $n-2$.

При $n \ge 30$ значимость r_S можно с достаточной точностью проверить на основании стандартного нормального распределения

$$z = [r_S] \cdot \sqrt{n-1}. \qquad (20.3)$$

Критические значения при $n \le 30$ приведены в таблице Е Приложения.

Предположим, например, что в выборке $n = 50$, $r_S = 0{,}30$. Проверим нулевую гипотезу $H_0 : \rho_S = 0$ против гипотезы $H_1 : \rho_S \ne 0$.

Пользуясь выражением (20.2), получим $\hat{t} = \dfrac{0{,}30}{\sqrt{\left(1 - 0{,}30^2\right)/48}} = 2{,}18$.

Полученное значение $\hat{t} = 2{,}18$ сравниваем с критическими значениями критерия Стьюдента с $\nu = 50 - 2 = 48$ степенями свободы на уровне значимости $\alpha = 0{,}05$: ${}_{0{,}975}t_{48} = -2{,}011$ и ${}_{0{,}025}t_{48} = 2{,}011$. Так как $\hat{t}$ попадает в критическую область, то нуль - гипотезу об отсутствии корреляционной связи между признаками X и Y следует отклонить на уровне значимости $\alpha = 0{,}05$.

Коэффициент ранговой корреляции Спирмена целесообразно применять в следующих случаях.

1. Когда необходимо быстро получить приближенную оценку коэффициента корреляции, а точный расчет очень громоздок.

2. Когда нужно перепроверить согласованность решений экспертов.

20.3. Коэффициент ранговой корреляции тау Кендалла. Все до сих пор рассмотренные нами коэффициенты корреляции допускают так или иначе разумное объяснение в терминах произведения моментов. Некоторые из них просто вытекают из формулы произведения моментов Пирсона, применённой к дихотомическим или порядковым данным, например, φ, r_s и r_{pb}. Другие, например, r_{tet} и r_{bis}, характеризуют попытки аппроксимировать коэффициент Пирсона. Английский статистик М. Кендалл (1955) предпринял попытку истолковать процесс измерения связей между переменными, не прибегая к принципу произведения моментов. Его усилия привели в конечном счете к возникновению нескольких, в сущности новых подходов к статистическому описанию.

В его системе, как и в случае с коэффициентом Спирмена, результаты наблюдения и для величины X, и для величины Y представляют собой n последовательных и не связанных между собой рангов. Сам же коэффициент корреляции строится на числе пар рангов, упорядоченных в одинаковом направлении как по величине X, так и по величине Y. Его мера, называемая «тау» и обозначаемая τ, представляет собой просто счётчик числа несовпадений в ранжировании X и Y.

Проиллюстрируем метод подсчёта коэффициента τ на конкретном примере.

Пусть восьми лицам присвоены ранги по величинам X и Y (табл. 20.1).

Таблица 20.3. Ранжирование по двум признакам

Лица	A	B	C	D	E	F	G	H
X	1	3	2	7	5	6	8	4
Y	3	2	1	4	7	8	6	5

Рассмотрим пару лиц, например, A и B. В ней A предшествует B по признаку X, а B предшествует A по признаку Y. Про такую пару говорят, что она образует инверсию (беспорядок). Инверсию образует и пара лиц (A, C). Пара же (A,D) инверсию не образует, так как A предшествует D и по признаку X, и по признаку Y. Среди 28 различных пар, которые

можно образовать из восьми элементов $\frac{n(n-1)}{2}$, инверсию образуют, кроме уже рассмотренных двух, пары (D,E), (D,F), (D,H), (E,G), (F,G). Таким образом, среди 28 пар инверсию образуют 7 пар. В остальных парах (21) сохраняется порядок, про них говорят, что они дают совпадение.

Обозначив число совпадений буквой P, число инверсий – буквой Q и общее число различных пар, которые образуют n элементов, – буквой N, выразим правило вычисления коэффициента тау Кендалла формулой: $\tau = \frac{P-Q}{N}$. Так как $N = \frac{n(n-1)}{2}$, то

$$\tau = \frac{2(P-Q)}{n(n-1)}. \qquad (20.4)$$

В рассмотренном выше примере $n = 8,\ P = 21,\ Q = 7,\ \tau = \frac{2(21-7)}{8(8-1)} = 0{,}5$.

Для упрощения подсчёта числа совпадений и инверсий заданные объекты упорядочиваются от 1 до n по X, как в таблице 20.4. Затем, начиная с первого объекта, подсчитывается, сколько раз его ранг по Y оказывается меньше, чем ранги объектов, расположенных правее. Полученное число записывается в строку, озаглавленную «Совпадение». В таблице 20.4 первый объект имеет ранг 3 по Y. Число 3 меньше, чем ранги пяти объектов, расположенных правее по Y, а именно 5, 7, 8, 4 и 6, поэтому в строке «Совпадение» для A записывается число 5. В строке «Инверсии» записывается число 2, показывающее, что правее A имеются всего два объекта, ранги которых меньше трёх. Для второго объекта (C) число совпадений, равное числу объектов, у которых ранг по Y больше 1, равно 6. Инверсий нет. Этот подсчёт продолжаем до конца таблицы. В последнем столбце записываются нули.

Таблица 20.4. Подсчет числа инверсии

Лица	A	C	B	H	E	F	D	G	Итого
X	1	2	3	4	5	6	7	8	
Y	3	1	2	5	7	8	4	6	
Совпадения	5	6	5	3	1	0	1	0	21
Инверсии	2	0	0	1	2	2	0	0	7

В порядке контроля: сумма совпадений и инверсий в каждом столбце должна быть равна числу столбцов, расположенных правее данного.

Значимость выборочного значения тау Кендалла проверяется с помощью одной из компонент, вычисляемой попутно с вычислением значения τ. Как было показано выше, для выборки объёма n тау Кендалла оп-

ределяется по формуле (20.4): $\tau = \frac{2(P-Q)}{n(n-1)}$. Если разность $P-Q$, т.е. разность между общим числом «совпадений» P и числом «инверсий» Q, в двух множествах рангов обозначить S, то $\tau = \frac{S}{n(n-1)/2}$.

Выборочное распределение S более удобно для исследования, чем τ. Когда n больше или равно 10, и X и Y в генеральной совокупности не коррелированны, то выборочное распределение S является приблизительно нормальным. В этом случае для проверки нулевой гипотезы о некоррелированности X и Y в генеральной совокупности применяется статистика

$$z = \frac{S^*}{\sqrt{n(n-1)(2n+5)/18}}, \qquad (20.5)$$

где $S^* = S+1$, если $S<0$ и $S^* = S-1$, если $S>0$.

Если нулевая гипотеза верна, то выборочное распределение для z хорошо описывается стандартным нормальным распределением.

Пример. Пусть $n=10$, а $S=9$, так что $\tau = \frac{9}{10\cdot 9/2} = 0{,}2$. Поскольку S положительно, то $S^* = S-1 = 9-1 = 8$ и величина $z = \frac{8}{\sqrt{10(10-1)(2\cdot 10+5)/18}} = 0{,}72$. Отсюда следует, что на уровне значимости $\alpha = 0{,}05$ нет оснований для отклонения нулевой гипотезы.

20.4. Квадрантная корреляция. Этот упрощенный критерий (Бломквист, 1951) позволяет проверить, имеется ли зависимость между двумя признаками X и Y, заданными в виде рядов измеренных значений. Вначале пары значений (x_i, y_i) отмечают в системе координат, которая делится значениями медиан $\tilde{x}$ и $\tilde{y}$ на 4 квадранта, так что каждая половина содержит точно одинаковое число пар значений. Если имеется нечетное число пар наблюдений, то горизонтальная медиана должна проходить через одну из точек, которая, таким образом, исключается. Затем подсчитывается число пар (точек), попавших в каждый квадрант. Очевидно, число точек, попавших в первый квадрант, будет равно числу точек, попавших в третий квадрант, а число точек, попавших во второй квадрант, равно числу точек, попавших в четвертый квадрант. В результате будет получено два числа, соответствующих числу точек в квадрантах. Их сравнивают с граничными точками (табл. М Приложения), определяемыми общим числом точек n и уровнем статистической значимости. Взаимозависимость параметров су-

ществует, если число пар значений в отдельных квадрантах достигает границ, указанных в таблице М Приложения или выходит за их пределы. Если выборки относятся к двумерному нормальному распределению, то этот критерий имеет асимптотическую эффективность по отношению к обычному коэффициенту корреляции 0,405.

Пример. Среди одних и тех же учащихся (16 человек) проведены две тестовые работы (по математике и по физике). В таблице 20.5 приведены баллы, набранные каждым учеником по математике и физике.

Таблица 20.5. Результаты тестирования учащихся по математике (X) и физике (Y). В последнем столбце значения медиан.

	A	B	C	D	E	I	j	F	G	H	K	L	M	N	O	P	медиана
X	51	41	50	40	29	44	41	34	72	54	18	26	53	64	47	54	46
Y	42	33	25	26	30	35	29	41	51	50	20	36	39	55	43	40	38

Предлагается проверить нуль-гипотезу об отсутствии зависимости между оценками.

В принципе строить для этого систему координат с упомянутыми выше медианами и наносить соответствующие точки не обязательно. Достаточно, пользуясь данными таблицы 20.5, заполнить соответствующие клетки таблицы 20.6.

Таблица 20.6. Распределение учащихся (пар значений) по четырем квадрантам, определяемым двумя медианами

		Число значений x	
		меньше медианы $\tilde{x}$ =46	больше медианы $\tilde{x}$ =46
Число значений y	Меньших медианы $\tilde{y}$ =38	a =1	b =7
	Больших медианы $\tilde{y}$ =38	c =7	d =1

Таким образом, статистика определяется парой чисел 1 и 7. Граничные точки критерия определяются по таблице М Приложения и на уровне статистической значимости $\alpha = 0{,}05$ равны 1 и 7. Откуда следует, что с 95%-ной надежностью можно отвергнуть нулевую гипотезу об отсутствии зависимости между рядами оценок.

20.5. Угловой критерий Олмстеда и Тьюки. Этот критерий хотя и похож на рассмотренный выше квадрантный критерий, однако требует большей информации. При его использовании уже недостаточно знать, сколько точек попадает в квадранты, образуемые пересечением медиан.

Признак учитывает и особенности расположения самих точек, степень их удаленности от медиан. Для его применения используется статистика S, которая вычисляется следующим образом.

1. Вначале n пар наблюдений (x_i, y_i), как и в рассмотренном квадрантном критерии, наносится в виде точек на плоскость, после чего образовавшееся корреляционное поле делится горизонтальной и вертикальной медианными линиями на четыре группы.

2. Точки в правом верхнем и левом нижнем квадрантах следует считать положительными, а остальные – отрицательными.

3. Начиная с правой стороны корреляционного поля, перемещают вертикальную прямую в направлении к вертикальной медиане с одновременным подсчетом точек и считают до тех пор, пока не встретится точка на другой стороне горизонтальной медианной линии. Сумме подсчитанных точек присваивается знак соответствующего квадранта. Аналогичный счет проводится в остальных квадрантах (с трех других сторон).

Сумма четырех полученных чисел и принимается в качестве соответствующей статистики. Обозначим ее буквой S (не путать со стандартным отклонением!).

Модуль значения статистики S сравнивается с критическими точками критерия S_α, которые приведены в таблице 20.7. При $|S| \geq S_\alpha$ предполагается корреляция, знак которой определяется знаком S.

Таблица 20.7. Критические точки углового критерия Олмстеда и Тьюки

α	0,10	0,05	0,02	0,01	0,005	0,002	0,001
S_α	9	11	13	14-15	15-17	17-19	18-21

1) При $\alpha \leq 0{,}01$ для малых n берется большая, для больших n – меньшая величина.

2) При $|S| \geq 2n - 6$ нужно критерий отбросить.

В качестве примера вновь обратимся к результатам тестирования учащихся по математике и физике, представленным в таблице 20.5.

Построив на координатной плоскости медианные прямые: $x = 46$ и $y = 38$, отметим на ней в строгом соответствии с данными, приведенными

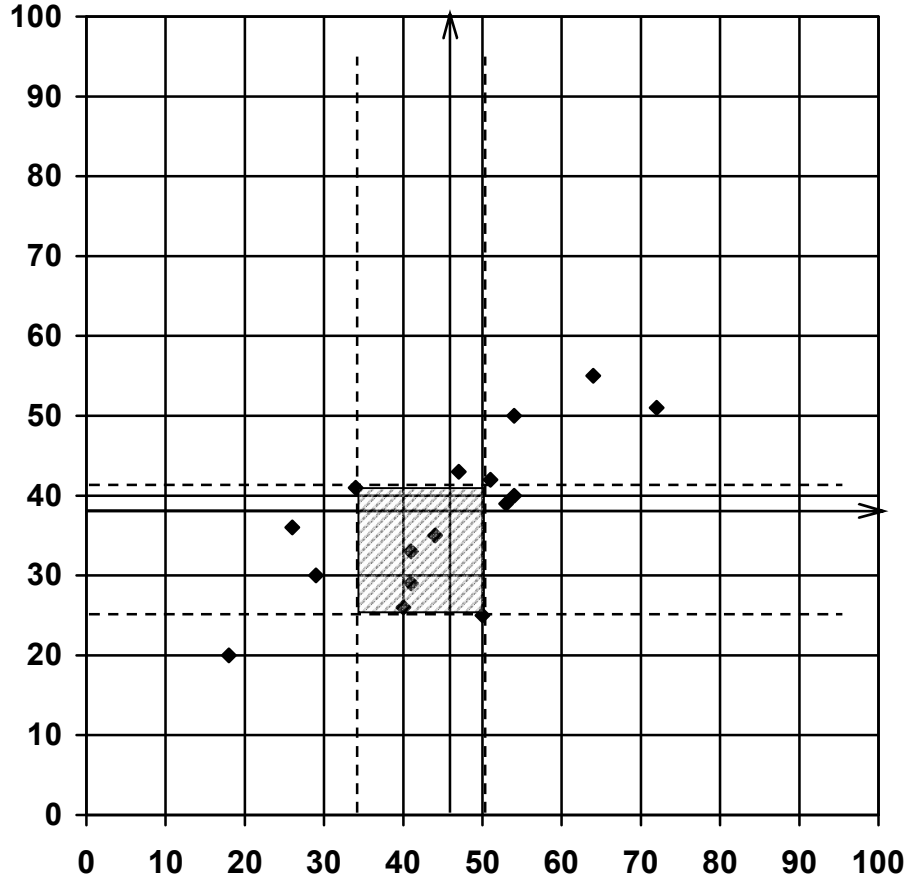


Рис.20.1. Расчет углового коэффициента

в таблице 20.5, все 16 точек. Пунктирными линиями изобразим положение прямых линий при перемещении которых в направлении к точке пересечения медиан, они впервые проходят через точки смежного квадранта. При перемещении вертикальной прямой справа налево до ближайшей пунктирной линии «заметаются» 6 точек первого квадранта, при перемещении горизонтальной прямой сверху вниз заметаются 5 точек первого квадранта, при перемещении вертикальной прямой слева направо – 3 точки третьего квадранта и, наконец, при перемещении горизонтальной прямой снизу вверх – 1 точка третьего квадранта. Всего 15 точек, причем все они из положительных квадрантов, поэтому S=+15. Так как $|S| < 2n - 6$, то угловой критерий можно использовать. Обращаясь к таблице 20.7 имеем $S_{0,01} = 15$. Так как $S \geq S_{0,01}$ и S>0, то имеется отчетливая положительная корреляция.

Глава 21. Частная и множественная корреляция и регрессия.

21.1. Частная регрессия. В главе 17 было показано, как можно по опытным данным найти зависимость одной переменной от другой, а именно как построить уравнение регрессии вида $\tilde{y} = a + bx$. При изучении влияния переменных x_1, x_2, ..., x_k на результирующий признак y необходимо строить регрессионные уравнения вида

$$\tilde{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k , \qquad (21.1)$$

в котором коэффициенты b_0, b_1, b_2,…, b_k называют коэффициентами регрессии.

В связи с уравнением (21.1) необходимо рассмотреть следующие вопросы:

а) как по эмпирическим данным вычислить коэффициенты регрессии b_0, b_1, b_2,…, b_k;

б) как интерпретировать эти коэффициенты;

в) как оценить тесноту связи между y и каждым из x_i в отдельности (при элиминировании действия остальных);

г) как оценить тесноту связи между y и всеми переменными x_1, x_2,…, x_k в совокупности.

Рассмотрим эти вопросы на примере построения двухфакторного регрессионного уравнения $\tilde{y} = b_0 + b_1 x_1 + b_2 x_2 ,$ (21.2)

Пример 1. Предположим, что изучается зависимость недельного бюджета свободного времени (y) от уровня образования (x_1) и возраста (x_2) определенной группы людей по данным выборочного обследования, результаты которого приведены в таблице 21.1.

Таблица 21.1. Результаты выборочного обследования: y – недельный бюджет свободного времени в часах, x_1 – уровень образования в эффективных (без второгодничества) годах обучения, включая среднее специальное или высшее образование, x_2 – возраст респондента в годах

Номер респондента	y	x_1	x_2
1	y_1	x_{11}	x_{21}
2	y_2	x_{12}	x_{22}
…	…	…	…
…	…	…	…
n	y_n	x_{1n}	x_{2n}
Среднее по столбцу	$\bar{y}$	$\bar{x}_1$	$\bar{x}_2$
Среднее квадратичное отклонение	S_y	S_1	S_2

При расчете коэффициентов уравнения множественной регрессии (21.2) полезно стандартизировать исходные, приведенные в таблице 21.1 эмпирические данные, пользуясь известными нам из пункта 6.4 формулами:

$$z_{1j} = \frac{x_{1j} - \bar{x}_1}{S_1}, \quad z_{2j} = \frac{x_{2j} - \bar{x}_{21}}{S_{21}}, \quad u_j = \frac{y_j - \bar{y}}{S_y}, \text{ где } \quad j=1, 2,\ldots, n.$$

В результате таких преобразований искомое регрессионное уравнение примет вид:

$$\tilde{u} = c_1 z_1 + c_2 z_2, \tag{21.3}$$

в котором стандартизированные коэффициенты регрессии c_1 и c_2 вычисляются по формулам:

$$c_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}, \quad c_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2}. \tag{21.4}$$

В этих формулах величины r_{1y}, r_{2y}, r_{12} представляют собой парные коэффициенты корреляции, вычисленные либо по исходным данным (таблица 21.1), либо по стандартизированным данным. Как следует из пункта 16.5, они могут различаться только знаком.

Предположим, что соответствующие парные коэффициенты имеют значения, приведенные в таблице 21.2.

Таблица 21.2. Парные значения корреляции между недельным бюджетом свободного времени y, уровнем образования (x_1) и возрастом респондента (x_2).

	y	x_1	x_2
	часов в неделю	лет обучения	возраст в годах
y	1	0,556	-0,131
x_1	0,556	1	-0,027
x_2	-0,131	-0,027	1
Среднее	31,6	9,0	30,2
Среднее квадратичное отклонение	16,5	2,9	11,5

Пользуясь этими данными, вычислим

$$\hat{c}_1 = \frac{0{,}556 - (-0{,}027)(-0{,}131)}{1 - (-0{,}027)^2} = 0Э55,$$

$$\hat{c}_2 = \frac{-0{,}131 - (-0{,}027) \cdot 0{,}556}{1 - (-0{,}027)^2} = -0{,}12.$$

В соответствии с этими данными уравнение регрессии примет вид $\tilde{u} = 0{,}55x_1 - 0{,}12x_2$.

Коэффициенты исходного регрессионного уравнения b_0, b_1, b_2 находятся по формулам:

$$b_1=c_1\left(\frac{S_y}{S_1}\right),\quad b_2=c_2\left(\frac{S_y}{S_2}\right),\quad b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2. \qquad (21.5)$$

Подставляя сюда данные из вышеприведенной таблицы, получим b_1=3,13; b_2=-0,17; b_0=8,50, откуда $\tilde{y}$=8,50+3,13x_1-0,17x_2.

В этом уравнении коэффициент b_1=3,13 показывает, что в среднем недельный бюджет свободного времени при увеличении уровня образования на один год и фиксированном значении возраста увеличивается на 3,13 часа. В то же время коэффициент $b_2 = -0{,}17$ показывает, что в среднем недельный бюджет сокращается на 0,17 часа при увеличении возраста на один год и при фиксированном уровне образования (x_1).

Коэффициент b_1 можно также рассматривать в качестве показателя тесноты связи между переменными y y и x_1 при постоянном значении x_2, а коэффициент b_2 – в качестве показателя тесноты связи между переменными y y и x_2, при постоянном значении x_1.

Аналогичную интерпретацию можно применить и к стандартизированным коэффициентам регрессии c_1 и c_2. Однако поскольку их значения вычисляются исходя из стандартизированных переменных, они являются безразмерными и в силу этого их можно использовать лишь при сравнении тесноты связи переменной y с переменными z_1 и z_2.

21.2. Частная корреляция. Рассмотренные в предыдущем пункте коэффициенты b_1 и b_2, (c_1 и c_2) измеряют одностороннее влияние (частную связь) одной случайной величины на другую при фиксированном значении третьей. Однако, приписанные им индексы при отсутствии соответствующих уравнений не позволяют установить о связи между какими величинами идет речь. Введем новую индексацию, устраняющую этот недостаток. Пусть, например, рассматривается связь между переменными x, y и z, тогда символом $b_{xy.z}$ будем обозначать коэффициент, измеряющий влияние случайной величины y на величину x при фиксированном значении величины z. Аналогично символом $b_{yx.z}$ будем обозначать коэффициент, измеряющий влияние случайной величины x на величину y при фиксированном значении z.

В этих условиях возникает необходимость иметь показатель, характеризующий связь в обоих направлениях. Таким показателем является *частный коэффициент корреляции*

$$r_{xy.z} = \sqrt{b_{xy.z} b_{yx.z}} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)}}. \qquad (21.6)$$

Частная корреляция выявляет зависимые переменные (по меньшей мере две) из независимых переменных. Точка в индексе $r_{xy \cdot z}$ отделяет две первые независимые переменные x и y от независимой переменной z.

Когда вместо букв x, y, z используются индексы 1, 2, 3, частную корреляцию между x_1 и x_2 при постоянном значении x_3 будем обозначать: $r_{12 \cdot 3} = \dfrac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$.

При циклической перестановке индексов, получим:

$$r_{13 \cdot 2} = \frac{r_{13} - r_{12} \cdot r_{23}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{32}^2\right)}},$$

$$r_{23 \cdot 1} = \frac{r_{13} - r_{12} \cdot r_{13}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{13}^2\right)}}.$$

В примере, рассмотренном в пункте 21.1, связь между случайными величинами y и x_1 при фиксированном значении x_2 равна:

$$\widehat{r}_{y12} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{\left(1 - r_{y2}^2\right)\left(1 - r_{12}^2\right)}} = \frac{0{,}556 - (-0{,}131)(-0{,}027)}{\sqrt{\left(1 - 0{,}131^2\right)\left(1 - 0{,}027^2\right)}} = 0{,}557.$$

В том же примере связь между случайными величинами y и x_2 при фиксированном значении x_1 равна:

$$\widehat{r}_{y2.1} = \frac{r_{y2} - r_{y1} r_{12}}{\sqrt{\left(1 - r_{y1}^2\right)\left(1 - r_{12}^2\right)}} = \frac{-0{,}131 - 0{,}556(-0{,}027)}{\sqrt{\left(1 - 0{,}556^2\right)\left(1 - 0{,}027^2\right)}} = -0{,}140.$$

Если имеются не три, а четыре переменные, то коэффициент частной корреляции между случайными величинами x_1 и x_2 при исключении влияния случайных величин x_3 и x_4 вычисляется по формуле:

$$r_{12.34} = \frac{r_{12.4} - r_{13.4} \cdot r_{23.4}}{\sqrt{\left(1 - r_{13.4}^2\right)\left(1 - r_{23.4}^2\right)}} = \frac{r_{12.3} - r_{14.3} \cdot r_{24.3}}{\sqrt{\left(1 - r_{14.3}^2\right)\left(1 - r_{24.3}^2\right)}}. \qquad (21.7)$$

Частный коэффициент корреляции проверяется на значимость так же, как и обычный коэффициент корреляции. Следует, однако, обратить внимание на то, что число степеней свободы при исключении каждой переменной уменьшается на единицу. Если исключается только одна пере-

менная, то число степеней свободы равно $m-2-1=m-3$. Вычисление частных коэффициентов корреляции обычно дает возможность элиминировать искажающее влияние тех факторов, которые в опыте или плохо контролируются или вообще не контролируются.

Расчет частных корреляций может внести ясность относительно взаимного влияния переменных при их неочевидной взаимозависимости. Если, например, корреляция между x_1 и x_2 основана только на общем влиянии x_3, то $r_{12.3} \approx 0$. Может случиться и так, что корреляция лишь поможет исключить мешающие переменные.

Пример 2. В одной из клиник была обстоятельно обследована выборка из 142 пожилых женщин. Три переменные– возраст (A), давление крови (B) и содержание холестерина в крови (C) имели следующие парные коэффициенты корреляции r_{AB}=0,3332, r_{AC}=0,5029, r_{BC}=0,2495.

Поскольку давление крови может быть связано с отложением холестерина на стенках сосудов, исследователям показалось интересным более детально изучить этот вопрос. Так как величины B и C увеличиваются с возрастом, возникает вопрос, можно ли отнести слабую связь между давлением крови и содержанием в ней холестерина лишь за счет возраста или же при каком-то возрасте существует более тесная связь. Влияние возраста исключается вычислением $r_{BC.A}$:

$$r_{BC.A} = \frac{r_{BC} - r_{AB} \cdot r_{AC}}{\sqrt{\left(1-r_{AB}^2\right)\left(1-r_{AC}^2\right)}},$$

$$\hat{r}_{BC.A} = \frac{0{,}2495 - 0{,}3332 \cdot 0{,}5029}{\sqrt{\left(1-0{,}3332^2\right)\left(1-0{,}5029^2\right)}} = 0.1005.$$

При переходе к t-критерию Стьюдента, получим

$$\hat{t}_{\text{выч}} = \frac{r}{\sqrt{\frac{1-r^2}{n-3}}} = \frac{0{,}1005}{\sqrt{\frac{0{,}9899}{139}}} = \frac{0{,}1005}{0{,}0844} = 1{,}191.$$

При числе степеней свободы, равном 142-2-1=139, на уровне статистической значимости $\alpha = 0{,}05$, $t_{\text{крит}}$=1,960. Так как $t_{\text{выч}}<t_{\text{крит}}$, то нет оснований отвергать нулевую гипотезу. Следовательно, связь между давлением крови и наличием в ней холестерина реализуется через возраст испытуемых. То есть сама по себе связь между давлением крови и наличием в ней холестерина статистически не значима.

Пример 3. Анализируя распространенность вредных привычек среди студентов КГПУ, были измерены значения парных коэффициентов корреляции между курением (A), употреблением пива (B), употреблением алкогольных напитков (C) и наркотиков (D). Их значения оказались равными: $r_{AB} = 0{,}28$, $r_{AC} = 0{,}34$, $r_{AD} = 0{,}28$, $r_{BC} = 0{,}40$, $r_{BB} = 0{,}17$, $r_{CD} = 0{,}30$. Так как обследованию было подвергнуто 1530 студентов, то, как легко проверить, все эти значения статистически значимы на уровне $\alpha = 0{,}01$. Вместе с тем следует заметить, что связь между употреблением наркотиков (D) и употреблением пива (B) является несколько сомнительной. Не исключено, что она опосредована двумя другими переменными (A и C). Чтобы выяснить это рассмотрим частную корреляцию между B и D при фиксированных параметрах A и C , т.е. $r_{BD.AC} = \dfrac{r_{BD\cdot C} - r_{AB\cdot C}\cdot r_{AD\cdot C}}{\sqrt{\left(1-r^2_{AB\cdot C}\right)\left(1-r^2_{AD\cdot C}\right)}}$.

Расчеты показывают, что $\hat{r}_{BD\cdot C} = \dfrac{0{,}17 - 0{,}40\cdot 0{,}30}{\sqrt{\left(1-0{,}40^2\right)\left(1-0{,}30^2\right)}} = 0{,}057$;

$\hat{r}_{AB\cdot C} = \dfrac{0{,}28 - 0{,}34\cdot 0{,}4}{\sqrt{\left(1-0{,}34^2\right)\left(1-0{,}4^2\right)}} = 0{,}167$; $\hat{r}_{AD\cdot C} = \dfrac{0{,}28 - 0{,}34\cdot 0{,}3}{\sqrt{\left(1-0{,}34^2\right)\left(1-0{,}3^2\right)}} = 0{,}1984.$

Используя полученные данные, получим:

$$r_{BD\cdot AC} = \frac{0{,}057 - 0{,}167\cdot 0{,}1984}{\sqrt{\left(1-0{,}167^2\right)\left(1-0{,}1984^2\right)}} = 0{,}0247 .$$

При переходе к t-критерию Стьюдента, получим

$$\hat{t}_{\text{выч}} = \frac{r}{\sqrt{\dfrac{1-r^2}{n-4}}} = \frac{0{,}0247}{\sqrt{\dfrac{0{,}99939}{1526}}} = \frac{0{,}0247}{0{,}0256} = 0{,}965 .$$

При числе степеней свободы, равном 1530-2-2=1526, даже на уровне статистической значимости $\alpha = 0{,}05$, $t_{\text{крит}}$=1,960. Так как $t_{\text{выч}} < t_{\text{крит}}$, то нет оснований отвергать нулевую гипотезу. Следовательно, связь между употреблением наркотиков и употреблением пива опосредована двумя другими переменными: курением и употреблением спиртного.

Между тем связь между курением и употреблением спиртного, при исключении двух других факторов (употребление пива и наркотиков), остается статистически значимой.

Действительно, как показывают расчеты,

$$\hat{r}_{AC\cdot BD} = \frac{r_{AC\cdot D} - r_{AB\cdot D}\cdot r_{CB\cdot D}}{\sqrt{\left(1-r^2_{AB\cdot D}\right)\left(1-r^2_{CB\cdot D}\right)}} = \frac{0{,}280 - 0{,}246\cdot 0{,}371}{\sqrt{\left(1-0{,}246^2\right)\left(1-0{,}371^2\right)}} = 0{,}2097 ,$$

откуда $\bar{t} = 8{,}4$, которое больше $t_{крит} = 2{,}576$.

Аналогично проверяются и другие связи.

21.3. Множественная корреляция (параметрический критерий). Для характеристики степени связи результирующего признака y служит множественный коэффициент корреляции.

Если возникает необходимость, характеризовать степень зависимости случайной величины x_1 одновременно от случайных величин x_2 и x_3, то она определяется *множественным коэффициентом корреляции* $R_{1.23}$:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}. \qquad (21.8)$$

В этом случае случайную величину x_1 называют выходной переменной (или зависимой переменной), а величины x_2 и x_3 входными переменными (или независимыми переменными). Точка в обозначении $R_{1.23}$ отделяет выходную переменную от двух входных переменных.

В результате циклической перестановки индексов можно получить формулы для $R_{2.13}$ и $R_{3.12}$. Значения множественного коэффициента корреляции лежат всегда между 0 и 1.

Обращаясь к примеру 1 (21.1), вычислим множественный коэффициент корреляции, выражающий зависимость y одновременно от x_1 и x_3.

Подставив в формулу $R_{y.12} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}}$ значения парных коэффициентов корреляции из таблицы 21.1, получим

$$R_{y.12} = \sqrt{\frac{0{,}556^2 + 0{,}131^2 - 2 \cdot 0{,}556 \cdot (-0{,}131)(-0{,}027)}{1 - 0{,}027^2}} = 0{,}5697.$$

Квадрат множественного коэффициента корреляции называется множественным *коэффициентом детерминации* $B=R^2$.

Значение $B=1$ означает, что выходная переменная точно определяется значениями входных переменных на основании множественной линейной регрессии (например, $\hat{y} = b_0 + b_1x_1 + b_2x_2$).

Можно показать, что квадрат коэффициента множественной корреляции $R_{1.23}$ связан с коэффициентом частной корреляции. $r_{13/2}$ следующим соотношением:

$$1-R^2_{1.23}=\left(1-r^2_{12}\right)\left(1-r^2_{13.2}\right). \qquad (21.9)$$

Действительно, подставив в выражение $r^2_{12}+\left(1-r^2_{12}\right)r^2_{13.2}$ значение $r_{13.2}=\dfrac{r_{13}-r_{12}\cdot r_{23}}{\sqrt{\left(1-r^2_{12}\right)\left(1-r^2_{23}\right)}}$ (см 21.6) и выполнив очевидные преобразования, получим $\dfrac{r^2_{12}+r^2_{13}-2r_{12}r_{13}r_{23}}{1-r^2_{23}}$, которое равно $R^2_{1.23}$ (см. 21.8). Таким образом, получаем, что $R^2_{1.23}=r^2_{12}+r^2_{13.2}\left(1-r^2_{12}\right)$, откуда и следует соотношение:

$$1-R^2_{1.23}=\left(1-r^2_{12}\right)\left(1-r^2_{13.2}\right). \qquad (21.9)$$

Аналогично выводится и соотношение

$$1-R^2_{1.234}=\left(1-r^2_{12}\right)\left(1-r^2_{13.2}\right)\left(1-r^2_{14.23}\right). \qquad (21.10)$$

Нуль-гипотеза, согласно которой параметр, соответствующий *R*, равен нулю (против: R>0), проверяется на основании F-критерия:

$\widehat{F}=\dfrac{R^2}{1-R^2}\cdot\dfrac{n-\left(k-u\right)-1}{k-u}$, $\quad \nu_1=k-u,\ \nu_2=n-\left(k-u\right)-1$, где $k-$ число случайных переменных; $u-$ число входных переменных, ранее названных независимыми переменными.

21.4. Множественная корреляция (непараметрический критерий). Множественный коэффициент корреляции W, иногда называемый коэффициентом конкордации, используется для измерения степени согласованности двух или нескольких рядов проранжированных значений переменных.

Коэффициент W вычисляется по формуле

$$W=\frac{12\cdot S}{k^2\cdot n\left(n^2-1\right)}\ , \qquad (21.11)$$

где k – число переменных; n – число индивидов, которые ранжируются;

$S=\sum_{i=1}^{n}\left(d_i-d\right)^2$, где d_i – сумма рангов по i-ой строке, а $d-$ среднее d_i.

Рассмотрим гипотетический пример. Восемь респондентов ранжированы по трём признакам: успеваемости, JQ и прилежанию (табл. 21.3).

Для вычисления множественного коэффициента корреляции W прежде всего вычислим среднее суммы рангов $d = 108/8 = 13{,}5$. После этого найдём $S = \sum_{i=1}^{8} (d_i - d)^2$, где d_i – сумма рангов, соответствующая i-му респонденту. Как видно из таблицы 21.3, эта сумма равна 328.

Таблица 21.3. Расчет множественного коэффициента корреляции

Респонденты	Успеваемость	JQ	Прилежание	Сумма рангов	$d_i - d$	$(d_i - d)^2$
1 – ый	2	3	1	6	- 7,5	56,25
2 – ой	8	8	7	23	9,5	90,25
3 – ий	1	1	2	4	- 9,5	90,25
4 – ый	4	4	5	13	- 0,5	0,25
5 – ый	7	6	8	21	7,5	56,25
6 – ой	3	2	4	9	- 4,5	20,25
7 – ой	6	5	6	17	3,5	12,25
8 – ой	5	7	3	15	1,5	2,25
Сумма	36	36	36	108		328

Таким образом, $W = \dfrac{12 \cdot 328}{3^2 \cdot 8 \cdot \left(8^2 - 1\right)} = 0{,}868$.

Значимость полученной величины W при $n > 7$ проверяется хи-квадрат критерием: $\chi^2 = \dfrac{12 \cdot S}{k \cdot n \cdot (n+1)}$ со степенями свободы $n - 1$. В рассмотренном примере $\chi^2 = \dfrac{12 \cdot 328}{3 \cdot 8 \cdot (8+1)} = 18{,}22$, степеней свободы $\nu = n - 1 = 8 - 1 = 7$. Для $\alpha = 0{,}05$ из таблицы С Приложения находим ${}_{0{,}025}\chi^2_7 = 16{,}01$. Поскольку наблюдаемое значение χ^2 больше критического значения, нулевую гипотезу о том, что не существует значимой связи между рассматриваемыми переменными (успеваемость, JQ и прилежание), нужно отвергнуть.

Глава 22. Элементы факторного анализа

22.1. Вводные замечания. Факторный анализ как особая область статистического анализа начал развиваться с начала XX столетия, прежде всего в психологии и по инициативе психологов. Однако вскоре он стал применяться и в других науках, таких, как социология, педагогика, экономика, биология, медицина, антропология и некоторых разделах физики.

Основной идеей факторного анализа, его главной задачей является концентрация информации, способность выражать большое число исходных косвенных признаков через меньшее число более ёмких внутренних характеристик явления, которые в то же время являются и наиболее существенными характеристиками исследуемого явления.

Потребность в такой концентрации информации впервые ощутили психологи, которые, анализируя психологические явления и процессы, сталкивались с многомерностью их описания, т.е. с необходимостью учитывать в анализе большое число показателей (параметров или признаков). Многие из рассматриваемых ими признаков были взаимосвязаны, в значительной мере дублируя друг друга, и лишь в косвенной форме отражали наиболее существенные, но неподдающиеся непосредственному наблюдению и измерению внутренние скрытые свойства явлений. Вскоре с похожими проблемами столкнулись представители многих других наук, имеющих дело с большим количеством непосредственно измеряемых признаков.

Именно в ситуациях подобного рода и оказалось естественным желание сконцентрировать информацию, выразить большое число исходных косвенных признаков через меньшее число более ёмких внутренних характеристик явления.

Можно сказать, что основная цель факторного анализа заключается в том, чтобы обнаружить скрытые общие факторы, объясняющие связи между наблюдаемыми признаками (параметрами) объекта.

Авторами основных концепций факторного анализа являются в первую очередь американские и английские ученые (Ч.Спирмэн, Л.Л. Тэрнстоун, Г.Х. Томсон, С.Л. Барт, Р.Б. Кеттелл и многие другие). С тех пор факторный анализ обогатился многими новыми методами и проник во многие новые области знания. Большую роль в этом процессе сыграло появление компьютеров и информационных технологий, на плечи которых легла огромная вычислительная работа, сопровождающая применение факторного анализа.

В настоящее время методы факторного анализа реализованы в форме информационных технологий, включенных практически в каждый пакет статистических программ, в частности в статистические пакеты STADIA, SPSS, STATISTICA, STATGRAPHICS, ЭВРИСТА, а также в некоторые пакеты математических программ.

Появился соблазн, не отягощая себя знакомством с теоретическими основами факторного анализа, применять его механически, следуя рекомендациям, сопровождающим программу. Однако этот путь представляется очень опасным. Формальное, без понимания смысла совершаемых действий применение методов факторного анализа может привести к серьезным ошибкам, понять значение которых или избежать которые исследователь будет не в состоянии.

В то же время рамки данного пособия не позволяют сколь-нибудь подробно и доказательно изложить теоретические основы факторного анализа. Поэтому мы, не прибегая к строгим математическим доказательствам, ограничимся разъяснением некоторых самых общих положений, знание которых облегчит исследователю использование методов факторного анализа.

Предполагая, что все расчеты, связанные с использованием факторного анализа, читатель будет выполнять на компьютере, математическое содержание этого метода опишем в самых общих чертах, зачастую существенно упрощая ситуацию. Несколько подробнее будет рассмотрен вопрос об интерпретации получаемых результатов.

22.2. Корреляционная матрица как исходный инструмент факторного анализа. Рассмотрим некоторую систему признаков, свойств, которыми в той или иной степени могут обладать или не обладать объекты некоторой совокупности и связь между которыми предполагается исследовать. Предположим, что каждое из этих свойств может быть «измерено» тем или иным способом, например с помощью тестов или анкетных опросов. Отобрав группу объектов (испытуемых, респондентов, школ, территориальных образований и т.п.) и подвергнув их соответствующим измерениям, оформим их результаты в виде таблицы, в верхней строке которой записаны сопоставляемые свойства (параметры), а в левом столбце – объекты измерения. Сжато эти данные можно представить в виде матрицы M, содержащей n столбцов, которым соответствуют упомянутые выше качества, свойства и N строк, каждой из которых соответствует свой объект (испытуемый).

$$\mathbf{M}=\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Nj} & \dots & x_{Nn} \end{pmatrix}.$$

В определенном смысле каждый признак можно было бы характеризовать (описать, «измерить») данными соответствующего столбца, однако для этого нужно приведенные в матрице данные стандартизировать по столбцам, для чего каждый элемент матрицы надо подвергнуть преобразованию: $z_{ij}=\dfrac{x_{ij}-\bar{x}_j}{S_j}$, где $\bar{x}_j$ выборочное среднее, а S_j – стандартное отклонение по элементам j-го столбца.

В результате матрица данных, обозначим её буквой А, примет вид:

$$A=\begin{array}{c} \begin{matrix} z(1) & z(2) & \dots & z(j) & \dots & z(n) \end{matrix} \\ \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1j} & \dots & z_{1n} \\ z_{21} & z_{22} & \dots & z_{2j} & \dots & z_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ z_{i1} & z_{i2} & \dots & z_{ij} & \dots & z_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ z_{N1} & z_{N2} & \dots & z_{Nj} & \dots & z_{Nn} \end{pmatrix} \end{array}.$$

Здесь $z(1), z(2), \dots, z(j), \dots, z(n)$ – качества (признаки), связь между которыми изучается в исследовании, а z_{ij} – число, характеризующее наличие или отсутствие качества $z(j)$ у i-го объекта (респондента).

При фиксированном значении j множество значений z_{ij}, $i=1, \dots, N$, то есть множество элементов j-го столбца, принято называть *стандартной формой задания качества (признака)* $z(j)$.

Для более детального изучения связей, существующих между признаками (свойствами), строят квадратную матрицу n-го порядка, элементами которой служат коэффициенты корреляций всех пар столбцов матрицы А, т.е. признаков $z(j)$.

$$R_1=\begin{array}{c} \\ z(1) \\ z(2) \\ \dots \\ z(i) \\ \dots \\ z(n) \end{array}\begin{array}{c} \begin{array}{cccccc} z(1) & z(2) & \dots & z(j) & \dots & z(n) \end{array} \\ \begin{pmatrix} 1 & r_{12} & \dots & r_{1j} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2j} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{i1} & r_{i2} & \dots & r_{ij} & \dots & r_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & r_{nj} & \dots & 1 \end{pmatrix} \end{array}.$$

Здесь r_{ij} – коэффициент корреляции между столбцами – свойствами $z(i)$ и $z(j)$. На диагонали этой матрицы записаны единицы, поскольку корреляция каждой переменной с самой собой равна 1. Такую матрицу называют *полной корреляционной матрицей* и обозначают R_1.

Часто на главной диагонали корреляционной матрицы вместо единиц ставятся квадраты коэффициентов множественной корреляции или другие оценки общности. Такие матрицы называют *редуцированными* и обозначают *R*.

Обратимся к полной корреляционной матрице R_1. Числа, стоящие в ее j-ом столбце характеризуют степень связи соответствующего признака (свойства) с каждым из остальных признаков (свойств).

22.3. Метод группировки признаков. При малом числе признаков (параметров) непосредственный визуальный анализ корреляционной матрицы оказывается достаточным для выделения из них отдельных групп высоко коррелированных между собой признаков. Первые работы по методам группировки параметров появились ещё до «изобретения» факторного анализа. Они предназначались для облегчения интерпретации эмпирически определяемой матрицы коэффициентов корреляции.

В числе первых выделяется метод П.В. Терентьева, получивший название *метода корреляционных плеяд*. Тереньтев предложил интерпретировать матрицу корреляций как полный взвешенный граф. Каждому такому графу можно поставить в соответствие последовательность обыкновенных графов. Именно задавшись некоторым порогом *h (h>0)*, заменим нулями все элементы матрицы коэффициентов корреляций, значения которых по модулю меньше заданного порога. Оставшиеся элементы заменим единицами. В итоге таких замен матрица коэффициентов корреляций преобразуется в матрицу, состоящую из нулей и единиц, по которой однозначно строится обыкновенный граф. Этот граф можно изобразить графически. Число его компонент связности (*корреляционных плеяд*) и сама то-

пология этих компонент в наглядной и обозримой форме характеризует все корреляционные связи между исходными параметрами, превышающими по модулю заданный порог. Систематический анализ такого рода, проведённый для достаточно большой последовательности убывающих по значению порогов: $h_1>h_2>h_3>\ldots>h_n$, даёт полное представление о структуре корреляционных связей между исходными параметрами.

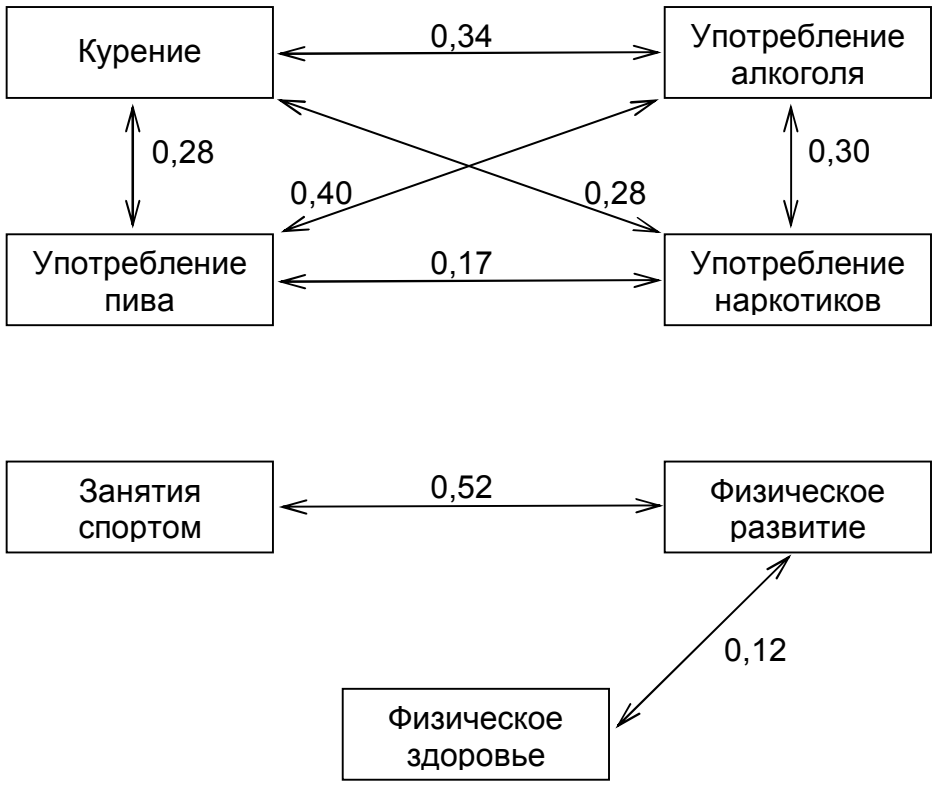
Метод корреляционных плеяд можно рассматривать как предварительную процедуру выделения групп сильно закоррелированных параметров. Особое значение этот метод приобретает в случае, когда получаемые на основе такой обработки группы оказываются связанными и по смыслу. В этом случае для каждой из выделенных групп можно построить один фактор, что, как уже отмечалось, сильно упрощает интерпретацию.

Появившаяся впервые в методе корреляционных плеяд интерпретация матрицы коэффициентов корреляций как некоторого *полного взвешенного графа* имела существенное значение для разработки других методов группировки параметров. Впоследствии эта интерпретация послужила основой для того, чтобы дать первые точные формулировки задач о группировке параметров. Это были вариационные задачи о выделении в полном взвешенном графе подграфов с наибольшими суммарными весами. Такая интерпретация матрицы коэффициентов корреляций имеет важное значение и для разрабатываемых в настоящее время общих методов анализа структуры произвольных матриц, определяющих связи в системах из большого числа элементов.

Рассмотрим в качестве примера связь между занятием спортом, физическим здоровьем, физическим развитием, курением, употреблением пива, алкогольных напитков и наркотиков среди студентов КГПУ по данным обследования 2003 года, выраженную таблицей 22.1.

Таблица 22.1. Полная корреляционная матрица седьмого порядка

		$z(1)$	$z(2)$	$z(3)$	$z(4)$	$z(5)$	$z(6)$	$z(7)$
Занятие спортом	$z(1)$	1,00	0,09	0,52	0,06	0,01	0,07	0,03
Физическое здоровье	$z(2)$	0,09	1,00	0,12	0,06	0,05	0,08	0,06
Физическое развитие	$z(3)$	0,52	0,12	1,00	0,07	0,02	0,12	0,07
Курение	$z(4)$	0,06	0,06	0,07	1,00	0,28	0,34	0,28
Употребление пива	$z(5)$	0,01	0,05	0,02	0,28	1,00	0,40	0,17
Употребление алкоголя	$z(6)$	0,07	0,08	0,12	0,34	0,40	1,00	0,30
Употребление наркотиков	$z(7)$	0,03	0,06	0,07	0,28	0,17	0,30	1,00



Выделяя наиболее значимые связи (коэффициент корреляции больше 0,15), можно построить две группы, два «графа», на которые распадается рассматриваемая система признаков.

Не вдаваясь в обсуждение причин отсутствия значимой корреляционной связи между физическим здоровьем и занятием спортом, заметим, что группа таких признаков, как курение, употребление пива, алкоголя и наркотиков, связана не только значимыми корреляционными связями, но и по содержательному смыслу рассматриваемых признаков.

Возникает вопрос о причинах такой связи. Как указывалось ранее, при рассмотрении вопроса о природе корреляционных связей, она может быть следствием того, что один из этих признаков, например, курение, является причиной трех остальных. Однако более правдоподобным является предположение, что существует некоторый гипотетический *фактор*, определяющий пристрастие к курению, пиву, алкоголю, наркотикам. Его содержательный смысл должен раскрываться специалистами в процессе интерпретации.

Если реализовать первое из этих предположений, рассматривая в качестве *факторов* такие признаки, как «курение» и «занятие спортом», то можно составить так называемую *факторную матрицу* (табл. 22.2), в которой все остальные качества выражаются значениями корреляции с ними. Эти значения корреляций, характеризующие влияние выделенного фактора

на остальные признаки называют *нагрузками соответствующих признаков (качеств)*.

Таблица 22.2. Нагрузки качеств «занятие спортом» и «курение»

		Занятие спортом	Курение
Физическое здоровье	$z(2)$	0,09	0,06
Физическое развитие	$z(3)$	0,52	0,07
Употребление пива	$z(5)$	0,01	0,28
Употребление алкоголя	$z(6)$	0,07	0,34
Употребление наркотиков	$z(7)$	0,03	0,28

Например, «физическое здоровье» характеризируется нагрузками 0,09 и 0,06, «Физическое развитие» – нагрузками 0,52 и 0,07. «Употребление пива» – соответственно нагрузками 0,01 и 0,28 и т.д. Важно заметить, что факторы разделили почти все признаки, кроме здоровья на две группы, в одну из которых вошли признаки, сильно связанные с первым фактором и практически не связанные со вторым, а во вторую – признаки, сильно связанные со вторым фактором и почти не связанные с первым.

Это описание качеств можно изобразить геометрически, приняв свойства $z(1)$ – «занятие спортом» и $z(4)$ – «курение» в качестве прямоугольной системы координат, а «нагрузки» каждого из остальных свойств в качестве их координат. В частности, рассматриваемый случай изображен на рисунке 22.1.

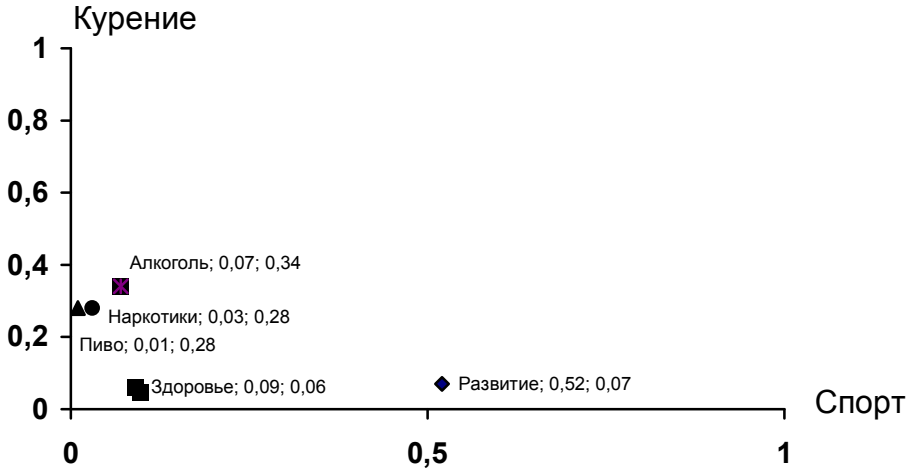


Рис. 22.1

К достоинствам такого изображения можно отнести то, что отчетливо видно, как отмеченные качества распадаются на две группы: первая включает физическое развитие, связанное с занятием спортом, а вторая – употребление спиртного, пива и наркотиков, связанное с курением. При большом числе переменных (качеств) можно выделить в качестве базовых

не две, а три переменных, изображая все остальные переменные в пространственной системе координат. Более того, можно использовать и большее число координат. К недостаткам следует отнести потерю информации. Во-первых, потеряны две очень важных переменных ($z(1)$ и $z(4)$), во-вторых, нет данных о степени взаимной связи отмеченных на рисунке переменных.

Во втором случае, когда факторы, обеспечивающие связи между исследуемыми переменными, ищутся «вне» этих переменных, возникает сложная проблема «измерения» степени влияния каждого из введенных, гипотетических факторов на каждый из рассматриваемых признаков, иначе говоря, *проблема вычисления факторных нагрузок.*

22.4. Выделение факторных нагрузок. Обратимся теперь к случаю, когда факторы выступают в качестве внешних по отношению к системе рассматриваемых признаков причин, определяющих связь между ними. Естественно, что число таких факторов должно быть существенно меньше числа рассматриваемых признаков. На практике исследователь подбирает их численность исходя из конкретных условий стоящих перед ним задач и получаемых результатов. Определившись с числом факторов, надо для каждого из них рассчитать факторные нагрузки. Естественно, что расчеты эти должны основываться на определенных принципах.

Предположим, что введено m гипотетических факторов: F_1, F_2, …, F_m. В соответствии с основной задачей факторного анализа, каждый признак $z(i),\ i = 1,2,...,n.$ должен линейно выражаться через эти факторы:

$$z(i) = a_{i1}F_1 + a_{i2}F_2 + ... + a_{im}F_m + d_iU_i, \qquad (22.1)$$

где $i = 1,2,...,n, \quad m < n$

Используемые в этом выражении факторы F_k, $k = 1,...,m$ называются *общими факторами*, так как они являются общими для всех признаков $z(i),\ i = 1,2,...,n$, а не для какого-то одного признака или группы признаков.

d_iU_i называется остатком (невязкой) или остаточным специфичным фактором.

Рассматриваемая зависимость может быть выражена таблицей 22.3.

Таблица 22.3. Общий вид таблицы факторных нагрузок

	Признаки (качества, свойства)						Факторы				Невязка
	$z(1)$	$z(2)$		$z(j)$		$z(n)$	F_1	F_2		F_m	$d_i U_{ii}$
$z(1)$	r_{11}	r_{12}		r_{1j}		r_{1n}	a_{11}	a_{12}		a_{1m}	$d_1 U_1$
$z(2)$	r_{21}	r_{22}		r_{2j}		r_{2n}	a_{21}	a_{22}		a_{2m}	$d_2 U_2$
…											
$z(i)$	r_{i1}	r_{i2}		r_{ij}		r_{in}	a_{i1}	a_{i2}		a_{im}	$d_i U_i$
…											
$z(n)$	r_{n1}	r_{n2}		r_{nj}		r_{nn}	a_{n1}	a_{n2}		a_{nm}	$d_n U_n$

Обычно в моделях факторного анализа предполагаются выполненными следующие предположения.

Общие факторы F_k, $k = 1,...,m$ являются либо случайными некоррелированными величинами с дисперсией 1, либо неизвестными неслучайными параметрами.

Остатки (остаточные факторы) U_i, $i = 1,2,...,n$ имеют нормальное распределение, не коррелированы между собой и не зависят от общих факторов.

В формуле (22.1) a_{ik} – неизвестные коэффициенты, которые и являются искомыми *факторными нагрузками*.

Естественно, что факторные нагрузки берутся не произвольно, а вычисляются по определенному правилу на основе либо полной, либо редуцированной корреляционной матрицы. Таких правил разработано много. Все они подчинены правилам оптимизации.

Если в качестве критерия оптимальности используют минимум расхождения между ковариационной матрицей исходных признаков и той, которая получается после оценивания нагрузок (мера «расхождения» двух матриц в данном случае есть евклидова норма их разности), то приходят к *методу главных компонент*.

Если критерием оптимальности является максимальная близость исходных корреляций признаков к тем, которые получены в модели после оценивания нагрузок, то возникает *анализ главных факторов*.

В качестве примера рассмотрим вычисление нагрузок двух первых общих факторов для матрицы (22.1) так называемым центроидным методом. Тем же методом могут быть вычислены факторные нагрузки третьего и всех последующих факторов.

Для построения первого фактора необходимо прежде всего редуцировать данную матрицу. В упрощенном варианте для этого в каждом столбце единицу заменим наибольшим элементом из числа оставшихся. Получим матрицу изображенную в таблице 22.4.

Таблица 22.4. Редуцированная корреляционная матрица

	Признаки (качества, свойства)								Факторы
	$z(1)$	$z(2)$	$z(3)$	$z(4)$	$z(5)$	$z(6)$	$z(7)$	$\sum r$	F_1
$z(1)$	0,52	0,09	0,52	0,06	0,01	0,07	0,03	1,30	0,433
$z(2)$	0,09	0,12	0,12	0,06	0,05	0,08	0,06	0,58	0,193
$z(3)$	0,52	0,12	0,52	0,07	0,02	0,12	0,07	1,44	0,480
$z(4)$	0,06	0,06	0,07	0,34	0,28	0,34	0,28	1,43	0,477
$z(5)$	0,01	0,05	0,02	0,28	0,40	0,40	0,17	1,33	0,443
$z(6)$	0,07	0,08	0,12	0,34	0,40	0,40	0,30	1,71	0,570
$z(7)$	0,03	0,06	0,07	0,28	0,17	0,30	0,30	1,21	0,403
$\sum r$	1,30	0,58	1,44	1,43	1,33	1,71	1,21	9,00	2,990

Вычисление нагрузок первого фактора выполняется по следующему простому правилу.

1) суммируются элементы каждой строки с учетом алгебраических знаков; сумма записывается в правом столбце $\sum r$;

2) складываются все суммы строк; получающаяся величина, обозначаемая буквой Т, в нашем случае равна 9,00;

3) суммы строк делят на $\sqrt{T}$, в нашем случае на 3. Полученные частные и будут искомыми нагрузками первого фактора. В таблице 22.3 они приведены в последнем столбце.

Для проверки правильности расчетов находят сумму всех факторных нагрузок, которая должна быть равна $\sqrt{T}$.

Вычисление нагрузок второго и последующих факторов выполняется сложнее и требует многих вспомогательных расчетов. Учитывая, что теперь все эти расчеты выполняются компьютерными программами, мы здесь приведем лишь готовую матрицу факторных нагрузок при двух факторах (табл. 22.5). Ее называют начальной, ибо редко когда поиски факторов (факторных нагрузок), осмысленно разделяющих анализируемые признаки, ею завершается. Характерен в этом отношении и рассмотренный нами пример. Хотя второй фактор достаточно отчетливо выделяет группу признаков, связанных с вредными привычками, нагрузки первого фактора мало о чем говорят. Анализ подобных ситуаций привел исследователей к мысли, что вызваны они, могут быть неудачным расположением осей, за-

висящим от метода, использованного для расчета факторов. На этой основе возникла идея вращения факторов.

Таблица 22.5. Матрица факторных нагрузок F_1 и F_2

	Признаки (качества, свойства)	Факторы	
		F_1	F_2
$z(1)$	Занятие спортом	0,433	- 0,536
$z(2)$	Физическое здоровье	0,193	- 0,080
$z(3)$	Физическое развитие	0,480	- 0,539
$z(4)$	Курение	0,477	0,301
$z(5)$	Употребление пива	0,443	0,340
$z(6)$	Употребление алкоголя	0,570	0,345
$z(7)$	Употребление наркотиков	0,403	0,243

22.5. Вращение факторов. Так как основная конфигурация векторов, соответствующих выделенным факторам, представляет собой неизменный элемент, вращающийся вокруг его начальной точки, принятой за начало координат, то ее проекции на различно расположенные оси могут взаимно преобразовываться, являясь в этом смысле эквивалентными.

Вращая таким способом систему координат, можно изменить набор факторных нагрузок. Эта операция в процедуре факторного анализа носит название *вращения*, и соответственно первый выбор «сырых» факторных нагрузок, полученных по окончании процесса *выделения* факторов каким либо методом, называется *исходным*. Аналогично можно говорить об исходной факторной матрице и о факторной матрице после поворота, а также о повернутых факторах.

Естественно, что существуют алгоритмы, позволяющие по величине поворота пересчитывать факторные нагрузки. Однако мы их здесь рассматривать не будем не только из-за сложности, но и потому, что они, как и правила выделения «сырых» факторов, заложены в соответствующих компьютерных программах.

Если в качестве примера подвергнуть такой обработке рассмотренную выше корреляционную матрицу (табл. 22.1), то получим диаграмму, изображенную на рисунке 22.2. Расчеты выполнены пакетом программ STATISTICA. Выделялись факторы методом главных компонент (Principal components), а вращались методом варимакс (Varimax).

Возникает вопрос, каков смысл всех этих преобразований факторов и факторных нагрузок? Почему нельзя остановиться на первом наборе факторных нагрузок, полученных по окончании процесса выделения факто-

ров? С математической точки зрения все возможные совокупности проекций, полученные в процессе вращения, эквивалентны. Их эквивалентность определяется тем, что каждую из них можно преобразовать в другую, сохраняя при этом их общность. Дело обстоит иначе при содержательной интерпретации факторных нагрузок (например, с психологической или социологической точки зрения).

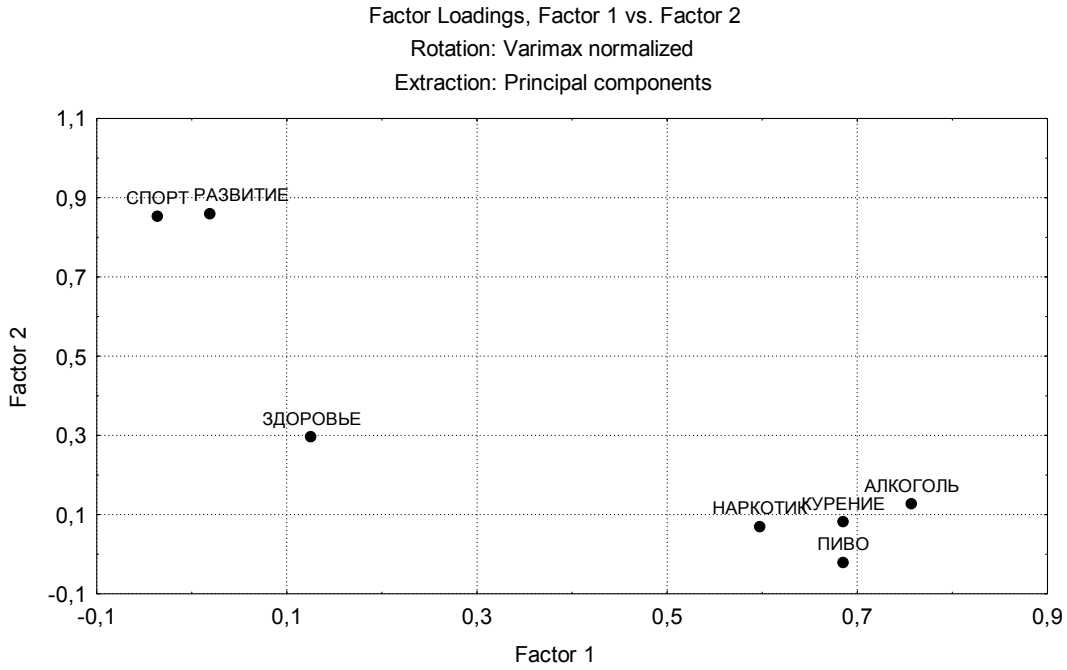


Рис. 22.2

В большинстве случаев существует какое-то одно определенное положение системы координат, дающее набор факторных нагрузок, имеющий особое значение. Психолог, социолог или экономист, применяющие метод факторного анализа, не могут довольствоваться результатами, наилучшими лишь с чисто математической точки зрения. Они должны стремиться к получению таких результатов, которые наилучшим образом соответствуют некоторой интерпретации факторных нагрузок, существенно связанной с проблематикой изучаемого явления. С этой точки зрения нельзя, как правило, остановиться на первом «сыром» наборе факторных нагрузок, а нужно использовать процесс вращения для нахождения такого положения системы координат, которое дает наиболее эффективные результаты. Нужно искать факторы, соответствующие каким-то существенным элементам, о которых уже есть некоторая информация и о которых с наибольшей вероятностью можно утверждать, что они отражают определенные реальные существующие в природе зависимости.

Исследователи, применяющие метод факторного анализа, считают, что существует одно положение осей координат, которое соответствует истинным факторам, а все другие возможные положения являются его математическими преобразованиями. То, что наряду с истинным положением системы координат существуют одновременно всевозможные ее случайные эквиваленты, есть просто неизбежный результат методов расчета факторных нагрузок. Этим и объясняется второй этап работы – вращение, в процессе которого нужно найти истинное положение системы координат, соответствующее реальным факторам.

Таким образом, следует так подбирать оси, чтобы результаты можно было проинтерпретировать в терминах предметной области. Это тонкое место факторного анализа, и ему следует уделять много времени при работе с реальными данными.

22.6. Работа с пакетом статистических программ Statistic. Прежде чем непосредственно обращаться к программам факторного анализа, необходимо определить совокупность сопоставляемых признаков. Это могут быть отдельные задания одного теста или разные варианты тестов; различные характеристики личности, особенности поведения и т.д. и т.п. Затем должна быть составлена таблица исходных данных, в которой по горизонтали, в верхней строке перечисляются исследуемые свойства, а в последующих строках приведены данные «измерений» по объектам. Эта таблица может быть оформлена в Microsoft Word или Excel и лишь затем перенесена в пакет статистических программ либо же изначально строится непосредственно в этом пакете.

Работа в блоке «Факторный анализ» пакета программ STATISTIC начинается с построения таблицы, содержащей необходимое число столбцов и строк, и занесения в нее подготовленных данных. Открыв стартовое окно Factor Analysis, в поле Input выберите Raw Data – Исходные данные. Из списка второго поля MD deletion выберите один из способов исключения пропущенных значений. Программой предусмотрено три способа такого исключения: Casewise, при котором вычеркиваются строки (случаи), в которых имеется хотя бы одно пропущенное значение; Pairwise, при котором игнорируются пропущенные случаи не для всех переменных, а лишь для выбранной пары, и, наконец, Mean Substitution, когда вместо пропущенных значений подставляются средние значения.

Щелкнув мышью по кнопке Variables, вы войдете в окно со списком переменных. В нем вы должны выделить переменные, с которыми предпо-

лагаете работать в дальнейшем. Нажав клавишу ОК, вернитесь в стартовое окно Factor Analysis.

Теперь все готово к реализации второго этапа – выделению факторов. Нажав кнопку ОК, вы попадете в окно **Define Method of Factor Extraction**. Здесь вам предоставляется возможность определить число выделяемых факторов (по умолчанию их предлагается 2) и метод их выделения.

Всего в программе 6 методов выделения факторов, объединенных общим заголовком **Extraction method – Метод выделения факторов:**

- Principal components – Метод главных компонент;
- Communalities=multiple;
- Iterated communalities – Итеративных общностей;
- Centroid method – Центроидный метод;
- Principal axis method – Метод главных осей;
- Principal axis method – Метод главных осей.

Выделив один из них, нажмите кнопку ОК. В результате появится окно **Factor Analysis Results – Результаты факторного анализа.**

В нем можно посмотреть факторные нагрузки выделенных факторов, для чего достаточно нажать кнопку Factor loadings.

Теперь можно перейти к третьему этапу – вращению факторов, для чего нажмите на кнопку Factor rotation. В результате появится окно **Factor Rotation**. В нем на выбор предлагается четыре метода вращения факторов:

- Varimax – варимакс;
- Biquartimax – биквартимах;
- Quartimax – квартимах;
- Equamax – эквимакс.

К каждому из этих методов приписан два термина: термин normalized, указывающий на то, что факторные нагрузки будут нормализованы, либо термин raw, указывающий на то, что факторные нагрузки не будут нормализованы.

Выделите одну из них, нажмите кнопку ОК. В результате получится таблица с численными значениями факторных нагрузок. Специальная кнопка Plot of loadings, 2D позволяет посмотреть двумерный график нагрузок.

Наряду с основной задачей программа позволяет решить все вопросы описательной статистики, проводить корреляционный и регрессионный анализ.