Основы корреляционного анализа. Анализ парных связей.

Анализ коэффициента корреляции.

Пример 1. В таблице приведен ряд, устанавливающий связь между уровнем IQ и уровнем средней успеваемости студентов по математике.

X – уровень IQ	75	85	90	100	105	110	110	115	115	120	125	130	140
Y – средняя успеваемость	3,1	3,1	3,5	3,7	3,8	4,0	4,2	4,3	4,6	4,7	4,8	4,9	5,0

Существует ли взаимосвязь между уровнем IQ (признак X) и средним уровнем успеваемости по математике (признак Y)?

Решение:

Представим исходные данные в расчетную таблицу.

Номер п/п	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	75	3,1	5625	9,61	232,5
2	85	3,1	7225	9,61	263,5
3	90	3,5	8100	12,25	315,0
4	100	3,7	10000	13,69	370
5	105	3,8	11025	14,44	399
6	110	4,0	12100	16,00	440
7	110	4,2	12100	17,64	462
8	115	4,3	13225	18,49	494,5
9	115	4,6	13225	21,16	529
10	120	4,7	14400	22,09	564
11	125	4,8	15625	23,04	600
12	130	4,9	16900	24,01	637
13	140	5,0	19600	25,00	700
Сумма:	**1420**	**53,7**	**159150**	**227,03**	**6006,5**

Вычислим выборочные средние:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1420}{13} = 109{,}23; \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{53{,}7}{13} = 4{,}13;$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i y_i) = \frac{6006{,}5}{13} = 462{,}04.$$

Теперь вычислим значения выборочных средних квадратических отклонений:

$$S_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - (\bar{x})^2} = \sqrt{\frac{1}{13}\cdot 159150 - 109{,}23^2} = 17{,}64;$$

$$S_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {y_i}^2 - (\bar{y})^2} = \sqrt{\frac{1}{13}\cdot 227{,}03 - 4{,}13^2} = 0{,}63.$$

Подставим в формулу:

$$r = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{S_x \cdot S_y} = \frac{462{,}04 - 109{,}23\cdot 4{,}13}{17{,}64\cdot 0{,}63} = \frac{10{,}9201}{11{,}1132} \approx 0{,}97.$$

Корреляционная связь между уровнем IQ и средним уровнем успеваемости по математике близка к линейной положительной. Чем выше уровень IQ у студентов, тем выше средний уровень успеваемости по математике, и наоборот.

Пример 2. Определить значимость выборочного коэффициента корреляции, вычисленного в примере 1.

Решение:

Выдвинем гипотезу $H_0: r = 0$ о том, что в генеральной совокупности отсутствует корреляция. Так как знак корреляции в результате решения примера 1 определен – корреляция положительна, то альтернативная гипотеза является односторонней вида $H_1: r > 0$.

Найдем эмпирическое значение t-критерия:

$$t_{\text{эмп}} = |r|\sqrt{\frac{n-2}{1-r^2}} = 0{,}97\sqrt{\frac{13-2}{1-0{,}97^2}} = 13{,}23.$$

Число степеней свободы равно $k = n - 2 = 13 - 2 = 11$, уровень значимости выберем равным $\alpha = 0{,}01$. По таблице прил. 2 находим критическое значение $t_{\text{крит}}(0{,}01; 11) = 3{,}11$.

Так как $t_{\text{эмп}} > t_{\text{крит}}$, то между уровнем IQ и средним уровнем успеваемости по математике существует статистически значимая корреляция.

Пример 3. По выборке $n = 100$ извлеченной из двумерной нормальной генеральной совокупности (X, Y), найден выборочный коэффициент корреляции $r = 0{,}2$. По уровню значимости 0,05 проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе H_1: $r \neq 0$.

Решение:

Найдем наблюдаемое (эмпирическое) значение критерия:

$$t_{\text{эмп}} = |r| \sqrt{\frac{n-2}{1-r^2}} = 0{,}2 \sqrt{\frac{100-2}{1-0{,}2^2}} = 2{,}02.$$

Конкурирующая гипотеза имеет вид $r \neq 0$, поэтому критическая область – двусторонняя.

По таблице прил. 2 при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $k = n - 2 = 98$ находим критическую точку двусторонней критической области $t_{\text{кр}}(0{,}05; 98) = 1{,}99$.

Так как $t_{\text{эмп}} > t_{\text{кр}}$ – отвергаем нулевую гипотезу о равенстве нулю генерального коэффициента корреляции. Другими словами, коэффициент корреляции значимо отличается от нуля.

Таким образом, X и Y коррелированны.

Пример 4. В таблице представлены результаты испытаний двух случайных величин X и Y $(n = 20)$.

i	x_i	y_i	i	x_i	y_i
1	4,08	2,14	11	4,31	6,29
2	6,91	3,00	12	2,34	5,52
3	7,42	1,73	13	3,82	3,11
4	3,58	4,24	14	3,98	5,70
5	5,16	3,27	15	3,24	2,60
6	5,19	2,83	16	2,88	5,13
7	4,10	4,22	17	6,19	1,44
8	5,37	4,40	18	5,86	2,20
9	5,02	2,19	19	2,67	3,58
10	6,19	3,20	20	4,36	3,90

Требуется определить выборочный коэффициент корреляции и проанализировать результаты.

Решение:

Вычислим необходимые для нахождения выборочного коэффициента корреляции оценки параметров распределения генеральной совокупности.

Номер п/п	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	4,08	2,14	16,6464	4,5796	8,7312
2	6,91	3,00	47,7481	9	20,73
3	7,42	1,73	55,0564	2,9929	12,8366
4	3,58	4,24	12,8164	17,9776	15,1792
5	5,16	3,27	26,6256	10,6929	16,8732
6	5,19	2,83	26,9361	8,0089	14,6877
7	4,10	4,22	16,81	17,8084	17,302
8	5,37	4,40	28,8369	19,36	23,628
9	5,02	2,19	25,2004	4,7961	10,9938
10	6,19	3,20	38,3161	10,24	19,808
11	4,31	6,29	18,5761	39,5641	27,1099
12	2,34	5,52	5,4756	30,4704	12,9168
13	3,82	3,11	14,5924	9,6721	11,8802
14	3,98	5,70	15,8404	32,49	22,686
15	3,24	2,60	10,4976	6,76	8,424
16	2,88	5,13	8,2944	26,3169	14,7744
17	6,19	1,44	38,3161	2,0736	8,9136
18	5,86	2,20	34,3396	4,84	12,892
19	2,67	3,58	7,1289	12,8164	9,5586
20	4,36	3,90	19,0096	15,21	17,004
Σ	92,67	70,69	467,0631	285,6699	306,9292

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{92,67}{20} = 4,6335; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{70,69}{20} = 3,5345;$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1}{20} \cdot 306,9292 = 15,3465.$$

Теперь вычислим значения выборочных средних квадратических отклонений:

$$S_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - (\bar{x})^2} = \sqrt{\frac{1}{20} \cdot 467,0631 - 4,6335^2} = 1,3725;$$

$$S_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} - (\bar{y})^2} = \sqrt{\frac{1}{20} \cdot 285,6699 - 15,3465^2} = 1,3381.$$

Подставим в формулу:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y} = \frac{15{,}3465 - 4{,}6335 \cdot 3{,}5345}{1{,}3725 \cdot 1{,}3381} \approx -0{,}5611.$$

Корреляционная связь между случайными величинами заметная.

Так как выборочный коэффициент корреляции отрицателен, то при возрастании одной случайной величины другая имеет тенденцию в среднем убывать.

Найдем наблюдаемое (эмпирическое) значение критерия:

$$t_{\text{эмп}} = |r| \sqrt{\frac{n-2}{1-r^2}} = 0{,}5611 \sqrt{\frac{20-2}{1-0{,}5611^2}} = 2{,}87.$$

Конкурирующая гипотеза имеет вид $r \neq 0$, поэтому критическая область – двусторонняя.

По таблице прил. 2 при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $k = n - 2 = 18$ находим критическую точку двусторонней критической области $t_{\text{кр}}(0{,}05; 18) = 2{,}10$.

Так как $t_{\text{эмп}} > t_{\text{кр}}$ – отвергаем нулевую гипотезу о равенстве нулю генерального коэффициента корреляции. Другими словами, коэффициент корреляции значимо отличается от нуля. Таким образом, с уверенностью 95 % можем полагать, что между рассматриваемыми числовыми совокупностями существует корреляционная связь.

Заметим, что наличие корреляционной связи можно утверждать даже при уровне значимости $\alpha = 0{,}02$, т. к. в этом случае $t_{\text{кр}}(0{,}02; 18) = 2{,}55$.

Пример 5. Определить достоверность взаимосвязи между показателями веса студента и максимального количества сгибания и разгибания рук в упоре лежа N у 10 исследуемых с помощью расчета рангового коэффициента корреляции, если данные выборок таковы:

85	75	73	77	77	81	78	90	83	80
26	20	25	22	27	28	16	15	18	24

Решение:

Составим вспомогательную таблицу, в которой результаты приведены в виде ранжированного ряда по весу студента.

Вес	85	75	73	77	77	81	78	90	83	80
d_x	**9**	**2**	**1**	**3,5**	**3,5**	**7**	**5**	**10**	**8**	**6**
N	26	20	25	22	27	28	16	15	18	24
d_y	**9**	**4**	**7**	**5**	**8**	**10**	**2**	**1**	**3**	**6**
$d_i = d_x - d_y$	0	–2	–6	–1,5	–4,5	–3	3	9	5	0

Заметим, т. к. четвертый и пятый испытуемый имеют равный вес, то им присваивается одинаковый ранг, равный 3,5.

$$\sum d_i^2 = 4 + 36 + 2{,}25 + 20{,}25 + 9 + 9 + 81 + 25 = 186{,}5.$$

Тогда выборочный коэффициент ранговой корреляции Спирмена

$$r_S = 1 - \frac{6 \cdot 186{,}5}{1000 - 10} = -0{,}13.$$

Так как $r_S = -0{,}13 < 0$, то между данными выборок наблюдается прямая отрицательная взаимосвязь, т. е. увеличение показателей веса вызывает снижение максимального количества сгибаний и разгибаний рук в упоре лежа в группе исследуемых.

Проверим значимость найденного рангового коэффициента корреляции. Найдем критические значения коэффициента ранговой корреляции Спирмена по таблице прил. 7 для $\alpha = 0{,}05$ и $n = 10$:

$$(r_S)_{\text{кр}} = 0{,}64.$$

Значение выборочного коэффициента ранговой корреляции $|r_S| = 0{,}13$ меньше значения $(r_S)_{\text{крит}} = 0{,}64$. Это говорит о том, что значение $r_S = -0{,}13$ не попало в область значимости коэффициента корреляции.

Таким образом, с вероятностью 95 % можно утверждать, что выявленная зависимость недостоверна.

Для уточнения результатов следовало бы повысить число исследуемых (увеличить объем выборки), а при отсутствии такой возможности высказанные оценки следует воспринимать с определенной осторожностью.

Итак:

1) увеличение показателей веса вызывает снижение максимального количества сгибаний и разгибаний рук в упоре лежа в группе исследуемых;

2) с уверенностью 95 % можно говорить о том, что выявленная зависимость недостоверна.

Пример 6. 14 жителям города *N* были заданы два вопроса: «Считаете ли Вы, что развитие космонавтики необходимо?» и соответственно «Согласились бы Вы предоставить себя в распоряжение ученым для научных экспериментов?» (с кодировками 0 = да и 1 = нет).

Результаты анкетирования представлены в таблице.

Вопрос		1		Итого
		0	1	
2	0	$a = 9$	$b = 5$	$a + b = 14$
	1	$c = 3$	$d = 11$	$c + d = 14$
	Итого:	$a + c = 12$	$b + d = 16$	$n = 28$

Оценить степень зависимости желания прогресса от личного здоровья.

Решение:

Чтобы воспользоваться формулой для вычисления коэффициента ассоциации, переведем исходные данные в таблицу сопряженности.

$$Q = \frac{ad - bc}{ad + bc} = \frac{9 \cdot 11 - 3 \cdot 5}{9 \cdot 11 + 3 \cdot 5} = \frac{84}{114} \approx 0{,}737;$$

$$\varphi = \frac{ad - bc}{\sqrt{(b + d)(a + c)(c + d)(a + b)}} = \frac{9 \cdot 11 - 3 \cdot 5}{\sqrt{16 \cdot 12 \cdot 14 \cdot 14}} = \frac{84}{194} \approx 0{,}433.$$

Можно утверждать, что в генеральной совокупности присутствует взаимная связь. Иными словами, существует значимая связь между желанием космических исследований и риском личного здоровья горожан.

Пример 7. 220 студентов были опрошены на предмет курения и употребления алкоголя.

Результаты анкетирования представлены в таблице.

Признак		X		Итого
		Пью	Не пью	
Y	Курю	$a = 40$	$b = 60$	$a + b = 100$
	Не курю	$c = 80$	$d = 40$	$c + d = 120$
	Итого:	$a + c = 120$	$b + d = 100$	$n = 220$

Оценить степень зависимости курения от употребления алкоголя.

Решение:

$$Q = \frac{ad - bc}{ad + bc} = \frac{40 \cdot 40 - 60 \cdot 80}{40 \cdot 40 + 60 \cdot 80} = -0{,}5;$$

$$\varphi = \frac{ad - bc}{\sqrt{(b + d)(a + c)(c + d)(a + b)}} = \frac{-3200}{12000} \approx -0{,}267.$$

Можно утверждать, что в генеральной совокупности отсутствует взаимная связь. Иными словами, не существует значимая связь между курением и употреблением алкоголя.

Пример 8. По результатам выборочных наблюдений были получены выборочные коэффициенты регрессии: $a_{yx} = -0{,}5$, $a_{xy} = -1{,}62$ Чему равен выборочный парный коэффициент корреляции.

Решение.

Коэффициент выборочной регрессии Y на X определяется по формуле $a_{xy} = r_в \frac{\sigma_y}{\sigma_x}$, а коэффициент выборочной регрессии X на Y определяется по формуле $a_{xy} = r_в \frac{\sigma_x}{\sigma_y}$. Выразим выборочный коэффициент корреляции $r_в$ из этих формул $r_в = a_{xy} \frac{\sigma_x}{\sigma_y}$, $r_в = a_{yx} \frac{\sigma_y}{\sigma_x}$. Перемножим эти два равенства, получим $r_в^{\,2} = a_{xy} \cdot a_{yx}$. Отсюда $r_в = \pm\sqrt{a_{xy} \cdot a_{yx}}$.Берем со знаком «-», если выборочные коэффициенты регрессии отрицательны, т.е. $a_{yx} < 0, a_{xy} < 0$, со знаком «+», если выборочные коэффициенты регрессии положительны. Подставим в полученную формулу значения выборочных коэффициентов регрессии $a_{yx} = -0{,}5$, $a_{xy} = -1{,}62$, получим

$r_в = -\sqrt{(-0{,}5) \cdot (-1{,}62)} = -\sqrt{8{,}1} = -0{,}9$.