Тема 3. Корреляция. Коэффициенты корреляции Пирсона и Спирмена

До сих пор статистические методы касались одной случайной переменной и ее распределения.

Однако многие проблемы в статистике касаются нескольких переменных. Во многих проблемах несколько переменных изучаются с целью установления их взаимосвязи или определения корреляции между ними.

Две случайные величины – X и Y – находятся в корреляционной зависимости, если каждому значению любой из них соответствует определенное распределение другой величины.

Чтобы определить корреляцию между двумя случайными величинами (X и Y), необходимо иметь две случайные выборки, одна из которых соответствует X, другая – Y.

Например, при анализе торговли ковровыми покрытиями было обнаружено изменение цен в % и соответствующие изменения в продаже. В результате имеем

X %	5	7	14	8	10	12	16	9
Y %	14	19	8	10	9	11	4	12

Взаимосвязь между этими случайными величинами можно проанализировать с использованием диаграммы рассеивания. С помощью этой диаграммы можно установить, есть ли связь между переменными и какого она вида. Для представленных данных диаграмма рассеивания имеет вид (рис. 2.18):

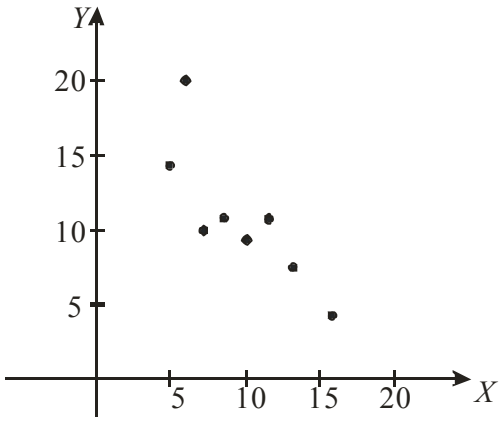
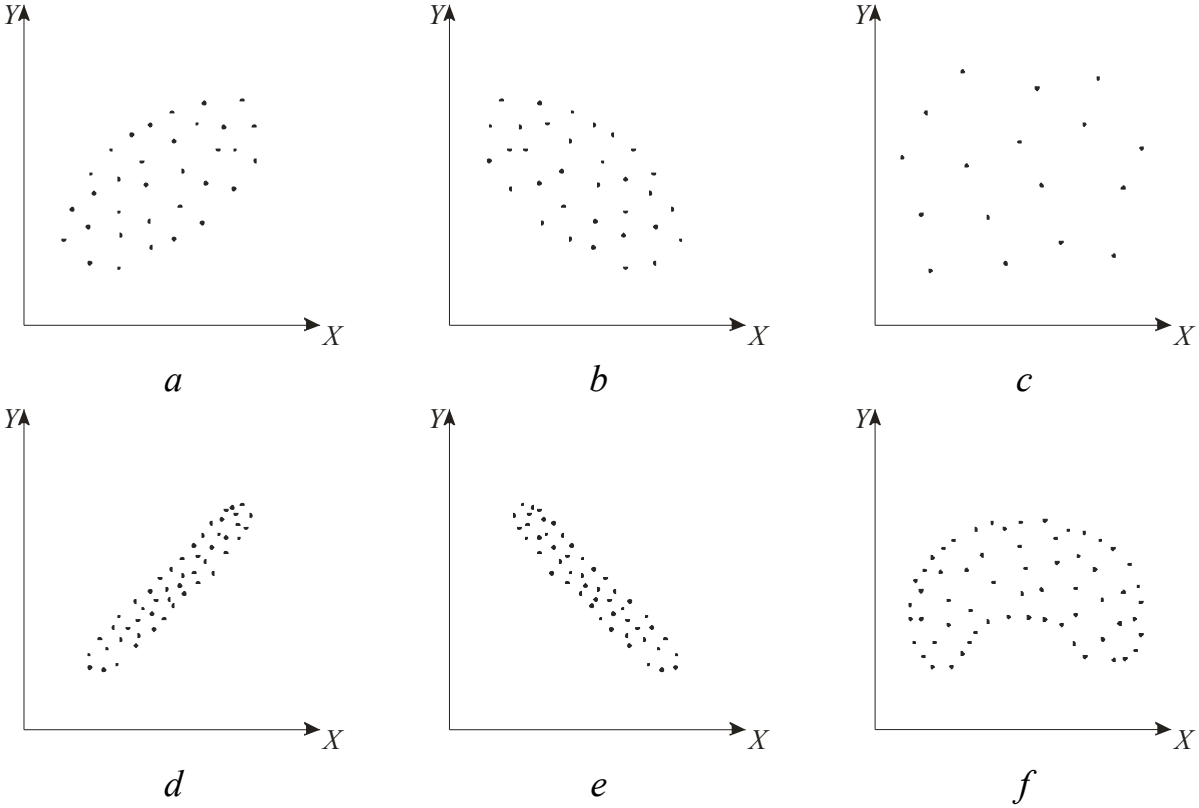


Рис. 2.18

Анализ этой диаграммы показывает, что при увеличении цен продажа имеет тенденцию к снижению. Более того, можно грубо оценить, что этот спад идет по прямой.

Взаимосвязь между переменными X и Y можно представить сле-

дующими диаграммами:



Взаимосвязь между X и Y, представленная на этих рисунках, классифицируется как:

a – слабая положительная линейная;

b – слабая отрицательная линейная (когда с возрастанием одной переменной другая убывает);

c – отсутствие связи;

d – сильная положительная линейная;

e – сильная отрицательная линейная;

f – нелинейная связь.

Для оценки линейной взаимосвязи между двумя случайными переменными X и Y используется *выборочный коэффициент корреляции Пирсона*

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{(n-1)S_x S_y}, \qquad (2.56)$$

где $\overline{X}$ – среднее арифметическое для X;

$\overline{Y}$ – среднее арифметическое для Y;

S_x – выборочное среднеквадратическое для X;

S_y – выборочное среднеквадратическое для Y;

n – объем выборок.

Коэффициент корреляции Пирсона предполагает, что случайные переменные X и Y являются непрерывного типа. Кроме того, предполагается, что они распределены по нормальному закону.

Это ограничивает применение коэффициента корреляции.

Существует *непараметрический аналог* коэффициента корреляции Пирсона – *ранговый коэффициент корреляции Спирмена*.

Коэффициент ранговой корреляции Спирмена находится по формуле

$$r = 1 - \frac{6\sum_{i=1}^{n}(r_i - S_i)^2}{n^3 - n}, \qquad (2.57)$$

где r_i и S_i – ранги i-го объекта по переменным X, Y;

n – число пар наблюдений.

То есть в данном случае проблема оценки тесноты связи решается с использованием ранжирования или упорядочивания объектов по степени выраженности измеряемых признаков. При этом каждому объекту присваивается определенный номер, называемый *рангом*.

Например. Объекту с наименьшим значением признака присваивается ранг 1, следующему за ним – ранг 2 и т. д.

При ранжировании иногда сталкиваются со случаями, когда величина проявления рассматриваемого признака одна и та же для нескольких объектов. В таких случаях объекты называются *связанными*. Связанным объектам приписываются одинаковые *средние ранги*.

Например. Если 4 объекта оказались равнозначными в отношении рассматриваемого признака и невозможно определить, какие из следующих рангов (4, 5, 6, 7) приписать этим объектам, то каждому объекту приписывается средний ранг, равный (4+5+6+7)/4 = 5.5.

При наличии связанных рангов ранговый коэффициент корреляции Спирмена вычисляется по формуле

$$r = 1 - \frac{\sum_{i=1}^{n}(r_i - S_i)^2}{1/6(n^3 - n) - (T_r + T_S)}, \qquad (2.58)$$

где $T_r = \frac{1}{12}\sum_{i=1}^{m_r}(t_r^3 - t_r)$; $T_S = \frac{1}{12}\sum_{i=1}^{m_S}(t_S^3 - t_S)$;

m_r, m_S – число групп неразличимых рангов у переменных X и Y;

t_r, t_S – число рангов, входящих в группу неразличимых рангов пе-

ременных X и Y.

Пример. Десять однородных предприятий были проранжированы по двум признакам – x_1 и x_2. В итоге имеем следующие выборки:

$$x_1 = (1;2.5;2.5;4.5;4.5;6.5;6.5;8;9.5;9.5);$$

$$x_2 = (1;2;4.5;4.5;4.5;4.5;8;8;8;10).$$

Определить коэффициент корреляции рангов.

Решение. В первой ранжировке имеем четыре группы неразличимых рангов. Во второй ранжировке имеем две таких группы:

$$r = 1 - \frac{\sum_{i=1}^{n}(r_i - S_i)^2}{1/6(n^3 - n) - (T_r + T_S)},$$

где $T_r = \frac{1}{12}\sum_{i=1}^{m_r}(t_r^3 - t_r)$; $T_S = \frac{1}{12}\sum_{i=1}^{m_S}(t_S^3 - t_S)$. В нашем случае $m_r = 4$, $m_S = 2$;

$t_r = 2$, $r = 1,2,3,4$; $t_s = 4$ для $s = 1$; $t_s = 3$ для $s = 2$. В результате

$$T_r = \frac{1}{12}\left[(2^3 - 2) + (2^3 - 2) + (2^3 - 2) + (2^3 - 2)\right] = \frac{24}{12} = 2;$$

$$T_s = \frac{1}{12}\left[(4^3 - 4) + (3^3 - 3)\right] = 7.42.$$

Используя формулу (2.57), имеем $r = 0.917$.

Примечание. Коэффициент корреляции рангов может использоваться для изучения связи между ординальными (порядковыми) переменными, которые еще называются *качественными*. В отличие от количественных переменных, для которых можно определить, на сколько или во сколько раз проявления одного признака у одного объекта больше (меньше), чем у другого, для качественных признаков этого определить нельзя.

Например. По некоторой дисциплине два студента имеют соответственно оценки «отлично» и «удовлетворительно». В этом случае можно утверждать, что уровень подготовки у первого студента выше, чем у другого, но нельзя сказать, на сколько или во сколько раз.

2.3.1. Свойства коэффициента корреляции

1. Коэффициент корреляции может принимать значения в интервале от –1 до +1 и равен +1 или –1 тогда, и только тогда, когда все точки

диаграммы лежат на прямой линии, т. е. в этом случае имеем функциональную зависимость.

2. Линейные преобразования, сводящиеся к изменению масштаба или начала отсчета случайных величин X и Y, не изменяют значения коэффициента корреляции

$$r(C_1X + C_2; C_3 + C_4) = z(X,Y).$$

3. Коэффициент корреляции между независимыми случайными величинами X и Y равен нулю. Обратное утверждение неверно, т. е. из равенства нулю коэффициента корреляции не следует независимость случайных величин X и Y. Если $r(X,Y) = 0$, то X и Y называются *некоррелированными*.

Только в одном случае некоррелированность случайных величин влечет их независимость. Это имеет место, если X и Y распределены по нормальному закону.

2.3.2. Значимость коэффициента корреляции

Выборочный коэффициент корреляции Пирсона является оценкой генерального коэффициента корреляции $\rho(X,Y)$.

В данном случае решается следующий вопрос. Может ли выборочный коэффициент корреляции случайно отличаться от нуля, а в действительности случайные переменные X и Y – некоррелированы?

Решение этого вопроса дается с помощью распределения вероятностей для выборочного коэффициента корреляции при условии, что генеральный коэффициент корреляции $\rho(X,Y) = 0$.

Существует таблица случайных отклонений от нуля произведения $|r_n| \cdot \sqrt{n-1}$ при условии, что $\rho(X,Y) = 0$ в зависимости от вероятности P и объема выборки n (табл. 2.12).

Таблица 2.12

Границы случайных отклонений значений $|r_n| \cdot \sqrt{n-1}$

n \ P	0.99	0.999	n \ P	0.99	0.999
10	2.29	2.62	25	2.47	3.03
11	2.32	2.68	30	2.49	3.07
12	2.35	2.73	35	2.50	3.10
13	2.37	2.77	40	2.51	3.13
14	2.39	2.81	45	2.52	3.15

15	2.40	2.85	50	2.53	3.16
16	2.41	2.87	60	2.536	3.184
17	2.42	2.90	70	2.541	3.198

Окончание табл. 2.12

n \ P	0.99	0.999	n \ P	0.99	0.999
18	2.43	2.92	80	2.546	3.209
19	2.44	2.94	90	2.550	3.219
20	2.45	2.96	100	2.553	3.226
			∞	2.576	3.291

Если выборочный коэффициент корреляции окажется больше приведенного в таблице граничного значения, то с надежностью P можно утверждать, что генеральный коэффициент корреляции $\rho(X,Y)$ отличен от нуля.

Значимость коэффициента корреляции можно проверить, решив следующую задачу проверки гипотез.

Выдвигаются гипотезы: $H_0 : \rho = 0$; $H_1 : \rho \neq 0$.

Задается уровень значимости α.

Статистика T определяется по формуле

$$T = r \cdot \sqrt{\frac{n-2}{(1-r^2)}}, \qquad (2.59)$$

где n – число пар данных.

Статистика T подчиняется t-распределению Стьюдента с $n-2$ числом степеней свободы. По таблице t-распределения определяется $t_{\alpha/2,\,n-2}$, $t_{1-\alpha/2,\,n-2}$. Если T, полученное по выборке, удовлетворяет условию $|T| > t_{1-\alpha/2,\,n-2}$, то H_0 отвергается и коэффициент корреляции считается значимым.

При проверке значимости коэффициента корреляции рангов исходят из того, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными, при $n > 10$, статистика

$$T = r \cdot \sqrt{\frac{n-2}{(1-r^2)}} \qquad (2.60)$$

имеет t-распределение Стьюдента с $k = n-2$ степенями свободы. Коэффициент корреляции значим на уровне α, если фактически наблю-

даемое значение t будет больше критического по абсолютной величине

$$|t| > t_{1-\alpha/2,\,n-2},$$

где $t_{1-\alpha/2,\,n-2}$ – табличное значение t-распределения Стьюдента при уровне значимости α и $k = n-2$.

При интерпретации коэффициента корреляции следует понимать, что:

– корреляция между двумя случайными величинами может быть вызвана влиянием других факторов, и для объяснения полученных результатов нужно хорошо знать область приложения;

– корреляция как формальное статистическое понятие не вскрывает причинного характера связи, т. е. нельзя указать, какую переменную принимать в качестве причины, а какую – в качестве следствия.

2.3.3. Задание для самостоятельной работы

1. В таблице представлены результаты измерения роста (x) и веса (y) у 12-ти студентов колледжа:

– представить диаграмму рассеивания;

– предположить значение коэффициента корреляции;

– вычислить коэффициент корреляции Пирсона.

x	65	73	70	68	66	69	75	70	64	72	65	71
y	124	184	161	164	140	154	210	164	126	172	133	150

Ответ: $r = 0.93$.

2. Какую интерпретацию можно дать коэффициенту корреляции между числом автомобильных аварий и возрастом водителей, который равен $r = -0.6$ (предполагается, что водители имеют по крайней мере по одной аварии).

3. Проверить гипотезу $\rho = 0$, если $n = 25$, $r = 0.35$.

Ответ: H_0 принимается, $\rho = 0$.

4. Проверить значимость коэффициента корреляции между переменными X и Y, значение которого $r = 0.740$ при уровне значимости $\alpha = 0.05$ и $n = 50$.

Ответ: H_0 отвергается.

5. Два эксперта проранжировали 10 предложенных им проектов реорганизации научно-производственного объединения (НПО) с точки

зрения их эффективности. Результаты представлены в виде

$$X_1 = 1,2,3,4,5,6,7,8,9,10; \quad X_2 = 2,3,1,4,6,5,9,7,8,10.$$

Ответ: $r = 0.915$