Министерство образования Российской Федерации

Тамбовский государственный технический университет

ИНФОРМАТИКА

(*корреляционный анализ и метод диаграмм рассеяния*)

Методические указания
по выполнению контрольных работ для студентов
заочного отделения и подготовке рефератов для студентов
дневного отделения специальностей 021100 и 002008

Тамбов
Издательство ТГТУ
2001

УДК 681.3 (075)
ББК з973.26я73-5
И 741

Утверждено Редакционно-издательским советом университета

Рецензент
профессор ТГУ им. Г. Р. Державина
В. А. Федоров

И 741 Информатика: Метод. указ. / Авт.-сост.: Ю. Л. Муромцев, Л. П. Орлова, Е. В. Бурцева, Д. Ю. Муромцев. Тамбов: Изд-во Тамб. гос. техн. ун-та, 2001. 36 с.

Методические указания содержат необходимые сведения по разделам: "Корреляционный анализ" и "Метод диаграмм рассеяния" по дисциплинам "Правовая информатика", "Информатика", "Анализ и синтез технических систем".

Предназначены для выполнения контрольных работ студентами заочной формы обучения и подготовки рефератов и курсовых работ студентами дневной формы специальностей 021100 и 002008.

УДК 681.3 (075)
ББК з973.26я73-5

введение

Важным этапом исследования систем и решения различных задач правовой деятельности является определение степени влияния одних переменных на другие с целью установления тесноты связей, выявления наиболее существенных, использования полученной информации при построении моделей и принятиях управленческих решений.

В настоящих указателях приводятся два метода выделения существенных связей между значительным числом переменных - метод диаграмм рассеяния и корреляционный анализ. Знание этих методов необходимо студентам дневной и заочной форм обучения специальностей 021100 и 002008 при изучении теоретических курсов "Правовая информатика", Информатика", "Анализ и синтез технических средств", подготовке рефератов и выполнении контрольных работ по данным курсам.

Данные методы выбраны не случайно. Один из них (диаграмм рассеяния) позволяет без трудоемких вычислений обработать большой массив статистической информации и наглядно представить результаты. Для применения второго метода (корреляционного анализа) имеются пакеты прикладных программ, позволяющие произвести необходимые вычисления на компьютере, например, Microsoft Excel 7.0.

Наряду с теоретическим материалом приводятся варианты заданий для контрольных работ студентам-заочникам и примеры выполнения работ.

1 Основные понятия и определения

При исследовании связей между различного рода явлениями важнейшей задачей является выделение факторов или признаков, которые оказывают основное влияние на изменение изучаемых явлений и процессов, т.е. вскрытие причинно-следственных отношений.

Признаки по их значению для изучения взаимосвязи делятся на:

1) факторные (факторы), они обуславливают изменение других, связанных с ними признаков;

2) результативные, они изменяются под действием факторов.

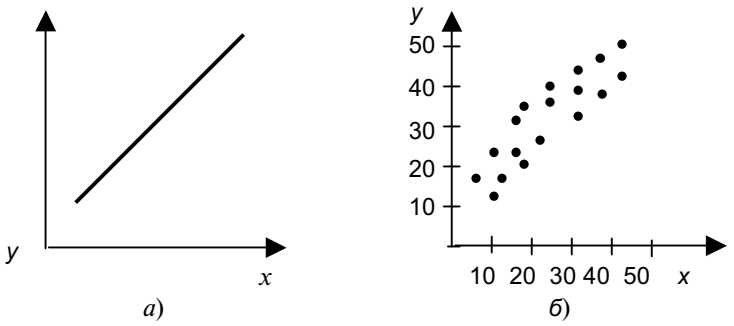
В статистике выделяют два вида зависимостей:

- *функциональная связь*, при которой определенному значению фактора x соответствует одно и только одно значение результативного признака y , эта связь проявляется во всех случаях наблюдения, и в каждом отдельном случае (см. рис. 1, *а*);

- *стохастическая связь* имеет место, когда зависимость y от x проявляется не в каждом отдельном случае, а в общем или среднем при большом числе наблюдений.

Корреляционная связь - частный случай стохастической связи, при котором изменение среднего значения y обусловлено изменением фактора x (см. рис. 1б).

Мощными и оперативными методами анализа связей между несколькими переменными исследуемых объектов являются *корреляционный анализ* и *метод диаграмм рассеяния*. Переменные в данном случае рассматриваются как система случайных величин $(x, y, z, \ldots)$, пара компонентов этой системы, например, x и y могут быть не связаны между собой (быть независимыми), находиться в линейной или нелинейной зависимо



а) *б*)

Рис. 1 – Функциональная (*а*) и

сти, причем эта зависимость может носить характер положительный или отрицательный.

Корреляция - статистическая зависимость между случайными величинами (СВ), не имеющими строго функциональной связи, при которой изменение одной из СВ приводит к изменению математического ожидания другой.

Корреляция (correlation, англ. - соотношение, соответствие) означает взаимосвязь между переменными (признаками), которая состоит в изменении средней величины одной из них в зависимости от значения другой.

Корреляционный анализ (КА) - математический аппарат в теории статистики для решения задач количественного определения тесноты связи между признаками (парная корреляция) и между результативными и несколькими факторными признаками (множественная корреляция). КА применяется, когда нельзя изолировать (исключить) влияние посторонних факторов либо потому, что они неизвестны, либо когда их изоляция невозможна.

Корреляционный анализ позволяет установить вид зависимости, ее характер и степень тесноты связи.

В корреляционном анализе выделяют *две основные задачи* исследования. *Первая* задача заключается в выявлении характера изменения результата в связи с изменением влияющего фактора при условии неизменности (однако наличии искажающего влияния) других факторов, т.е. здесь определяется форма связи. *Вторая* задача состоит в определении степени влияния искажающих факторов, т.е. оценке тесноты связи.

Корреляционно-регрессионный анализ включает методы определения тесноты и направления связи (корреляция) и установление аналитического выражения (формы) связи (регрессия).

Корреляционное поле представляет собой график в прямоугольных осях координат, заполненный точками, характеризующими каждую статистическую единицу (x_i, y_i) по двум коррелируемым признакам: факторный признак x откладывается по масштабной оси абсцисс, а результативный y - по масштабной шкале оси ординат (см. рис. 1, *б*). Оно в какой-то степени отражает информацию корреляционной таблицы, в которой записываются частоты сочетаний значений двух взаимосвязанных величин $(x\,y)$. Значениям одной величины (x) отводятся столбцы, другой (y) - строки. В клетках, образуемых пересечением столбцов и строк, записывается число случаев, в которых одни значения сочетаются с другими.

По корреляционному полю или корреляционной таблице делаются некоторые предварительные выводы, например, о прямой или обратной связи, о тесноте связи и др.

Корреляционная зависимость есть взаимосвязь между признаками, состоящая в том, что средняя величина $(\bar{y})$ значений одного признака y меняется в зависимости от изменения другого признака (x).

Метод диаграмм рассеяния выделился из метода случайного баланса применительно к условиям пассивного эксперимента. Диаграмма рассеяния (ДР) представляет собой особое графическое изображение значений экспериментальных данных о входных переменных x_i, $i = 1, 2, \ldots, n$ и выходной переменной *y*, которое компактно характеризует связь между ними. В некотором смысле ДР можно рассматривать как упрощение изображения корреляционного поля.

Пример 1 Пусть собраны статистические данные о переменных $x_1, x_2, \ldots, x_5, y$ (табл. 1).

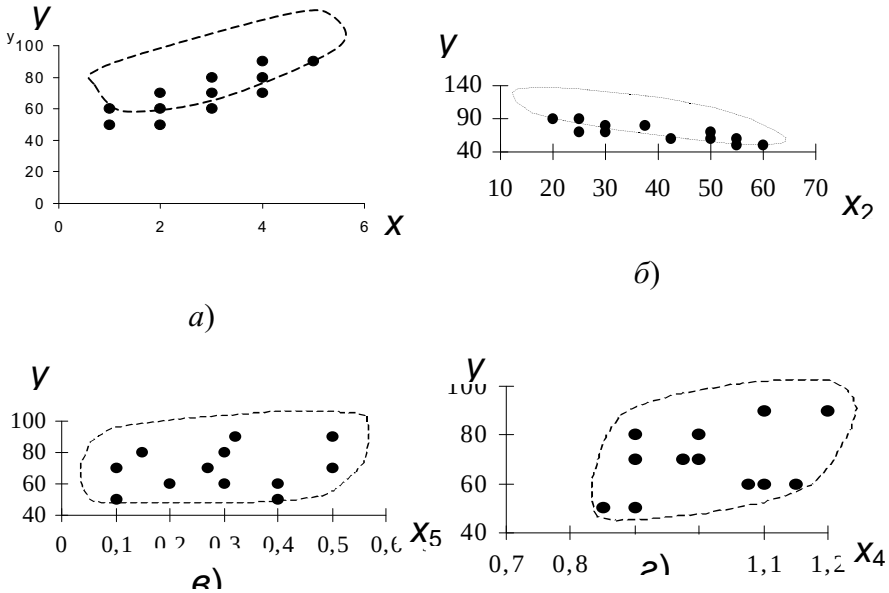
Таблица 1

№	x_1	x_2	x_3	x_4	x_5	y
1	2	50	0,2	1,1	5	60
2	2	25	0,1	1	9	70
3	1	42,5	0,3	1,15	12	60
4	3	30	0,27	0,9	5	70
5	3	37,5	0,3	1	9	80
6	4	30	0,15	0,9	7	80
7	1	55	0,4	0,9	3	50
8	4	50	0,5	0,975	12	70

9	4	25	0,5	1,1	6,5	90
10	2	60	0,1	0,85	2	50
11	3	55	0,4	1,075	11	60
12	5	20	0,32	1,2	7,5	90
$\bar{x}_i$	~3	40	~0,3	~1	7,4	
Δx_i	0,5	2,5	0,1	0,02	0,3	

На рис. 2, *а* - *г* приведены корреляционные поля для пар (y, x_1), (y, x_2) и т.д., на рис. 3 показан пример криволинейной зависимости (y, x_5), а на рис. 4 - построенные диаграммы рассеяния.

Корреляционные поля строятся простым обозначением точек, соответствующих наблюдениям $(x_{ij},\ y_j),\ j = 1, 2, \ldots, N$. Так значение $(x_{11} = 2,\ \ y_1 = 60)$ из первой строки данных в табл. 1 выделено на графике рис. 2, *а*. Из корреляционного поля (x_1, y) (рис. 2, *а*) видно, что значение y при увеличении x_1 в среднем возрастает, а при увеличении x_2 (см. рис. 2, *б*) - убывает. Изменение переменных x_3 и x_4 (см. рис. 2, *в*, *г*) слабо сказывается на величине y. На рис. 3 показан пример криволинейной зависимости y от x_5.



а) *б*) *в*) *г*)

Рис. 2 Корреляционные поля

Методику построения ДР рассмотрим на примере пары (x_1, y). Сначала определяется среднее значение $\bar{x}_1$, высокая точность здесь не требуется. В нашем случае $\bar{x}_1$ примерно равно 3.

Далее в окрестности среднего значения $\bar{x}_1$ выделяется "центральная" или промежуточная зона $[\bar{x}_1 - \Delta x_1;\ \bar{x}_1 + \Delta x_1]$. Величина Δx_1 пропорциональна ошибке в определении x_1.

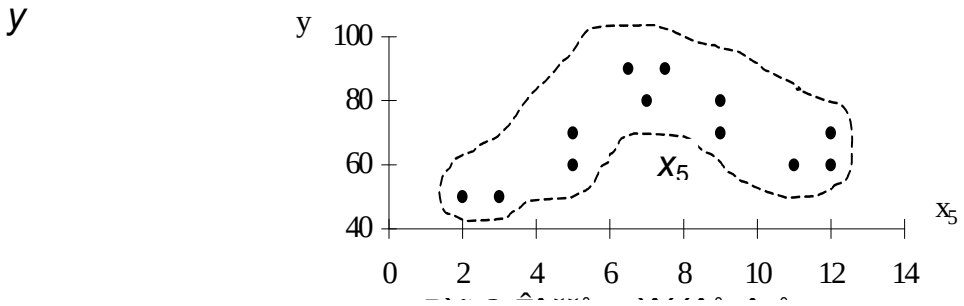


Рис. 3 Корреляционное поле (криволинейная зависимость)

Значения наблюдений (y_j, x_{1j}), $j = 1, 2, \ldots,$ для которых x_{1j} - попадает в центральную зону, из дальнейшего рассмотрения при построении ДР исключаются.

Пусть $\Delta x_1 = 0{,}5$, тогда центральная зона есть интервал $[2{,}5;\ 3{,}5]$ и наблюдения (строки) под номерами 4, 5, 11 (для них $x_1 = 3$) не участвуют в дальнейшем построении.

Затем все значения y при $x_1 < 2{,}5$ откладывались слева полосы над x_1 (над знаком "-"), а значения y при $x_1 > 3{,}5$ - справа полосы (над знаком "+"). Аналогично строится ДР для (x_2, y). Полоса над x_i соответствует "центральной" зоне значений x_{ij}, $j = 1, \ldots, N$.

Как видно из сопоставления корреляционного поля и ДР для пары (x_1, y), если с увеличением x_1 значение y в среднем возрастает, т.е. имеет место положительная корреляция, то на ДР правая группа точек смещена вверх по сравнению с левой. При отрицательной корреляции для x_2 правая группа точек смещена вниз.

2 МЕТОД ДИАГРАММ РАССЕЯНИЯ

Метод ДР удобно применять при большом числе входных переменных x_i, $i = 1, 2, \ldots, n$ на начальном этапе анализа связей между x_i и выходной переменной y, т.е. до корреляционного анализа. ДР позволяют более компактно графически отобразить связи между y и переменными x_i. Степень влияния x_i на y оценивается двумя показателями - величиной вклада B_i и числом выделившихся точек W_i.

Для определения вклада сначала находятся медианные значения $Me_{i(-)}$, $Me_{i(+)}$ для левой и правой совокупности точек ДР. При расчете медианы некоторой совокупности значений y_j, предварительно эти значения записываются в виде ранжированного ряда, т.е. в порядке возрастания (или убывания). Пусть имеется ранжированный ряд $y_1, y_2, \ldots, y_n$, тогда

$$Me = \begin{cases} y_{(M+1)/2}, & \text{если M - нечетное;} \\ \dfrac{1}{2}\left(y_{M/2} + y_{M/2+1} \right), & \text{если M - четное,.} \end{cases} \qquad (1)$$

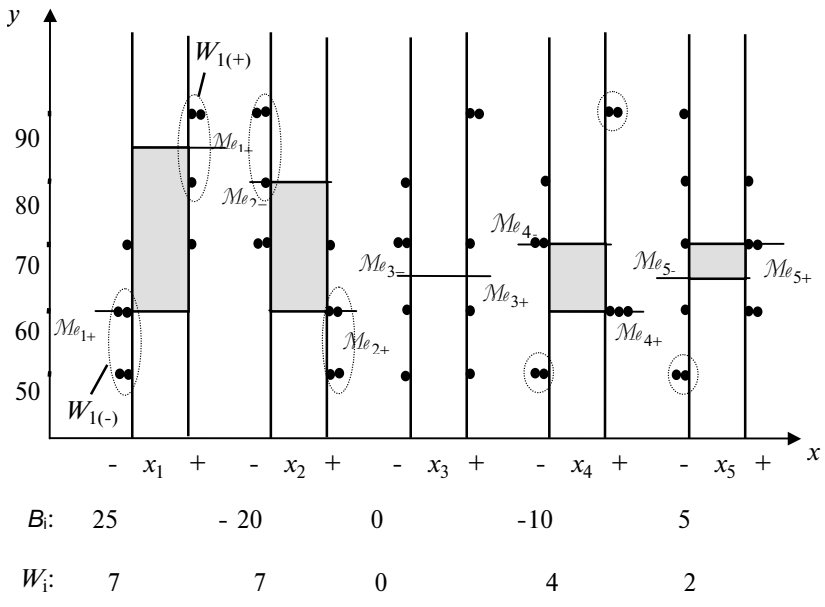


Рис. 4 Диаграммы рассеяния, вклады B_i и числа выделившихся точек W_i

Рассчитав по формуле (1) значения медиан $Me_{i(-)}$, $Me_{i(+)}$ оценивают величину вклада как разность медиан

$$B_i = Me_{i(+)} - Me_{i(-)}. \qquad (2)$$

Определение 1 Вкладом B_i, характеризующим степень влияния изменения входной переменной x_i на выходную переменную y, называют разность медианных значений $Me_{i(-)}$, $Me_{i(+)}$, полученных соответственно для правой и левой совокупностей точек диаграммы рассеяния (y, x_i).

Чем больше абсолютное значение вклада $|B_i|$, тем сильнее влияние x_i на y. Знак вклада определяет направление этого влияния: при $B_i > 0$ с увеличением x_i значение y в среднем возрастает, в случае $B_i < 0$ с увеличением x_i значение y уменьшается.

Пример 1 Для условий примера 1 $Me_{1(-)}$ находится для левой совокупности точек, которые в порядке возрастания можно записать

$$y_1 = 50;\ y_2 = 50;\ y_3 = 60;\ y_4 = 60,;\ y_5 = y_M = 70.$$

Так как $M = 5$ нечетное, то по формуле (1)

$$Me_{1(-)} = y_{(M+1)/2} = y_3 = 60.$$

Для правой совокупности точек

$$y_1 = 70;\ \ y_2 = 80;\ \ y_3 = 90;\ \ y_4 = y_M = 90$$

медиана равна

$$Me = \frac{1}{2}\left(y_{M/2} + y_{M/2+1}\right) = \frac{1}{2}(y_2 + y_3) = \frac{1}{2}(80 + 90) = 85.$$

Таким образом, используя формулу (2), получаем значение вклада для переменной x_1

$$B_1 = Me_{1(+)} - Me_{1(-)} = 85 - 60 = 25.$$

Величине B_1 соответствует заштрихованный столбик (см. рис. 4).

Значения медиан и вкладов показаны на рис. 4 под диаграммами рассеяния.

Определение 2 Выделившимися точками слева $W_{i(-)}$ и справа $W_{i(+)}$ для ДР определяются следующим образом (y, x_i):

1) если левая медиана ниже правой $(Me_{i(-)} < Me_{i(+)})$, то $W_{i(-)}$ образуют точки левой совокупности, находящиеся ниже наименьшего значения точек правой совокупности, а $W_{i(+)}$ образуют точки справа, расположенные выше наибольшего значения точек слева;

2) если левая медиана выше правой $(Me_{i(-)} > Me_{i(+)})$, то $W_{i(-)}$ образуют точки левой совокупности, находящиеся выше наибольшего значения точек справа, а $W_{i(+)}$ - точки правой совокупности, расположенные ниже наименьшего значения точек слева.

Общее число выделившихся точек для ДР (y, x_i) равно

$$W_i = W_{i(-)} + W_{i(+)}. \qquad (3)$$

Чем больше W_i, тем сильнее влияние x_i на y. Показатель W_i обычно считается более важным, чем вклад B_i при выделении существенных связей между x_i, $i = \overline{1,n}$ и y.

Замечание 1 Если $Me_{i(-)} = Me_{i(+)}$ или разность между медианами мала, то выделившиеся точки определяются при наличии явного смещения между интервалами значений левой и правой совокупности точек (см. рис. 4 для x_3 и x_4).

Пример 1, *б* Для условий примера 1 у ДР (y, x_1) левая медиана $Me_{1(-)}$ ниже правой $Me_{1(+)}$. Поэтому $W_{1(-)}$ образуют точки (50, 50, 60, 60), которые расположены ниже наименьшего значения справа - 70, т.е. $W_{1(-)} = 4$.

Справа выделившимися точками являются (80, 90, 90), которые выше наибольшего значения слева - 70, т.е. $W_{1(+)} = 3$. Таким образом, в соответствии с формулой (3)

$$W_1 = W_{1(-)} + W_{1(+)} = 7.$$

Аналогично для ДР (y, x_2):

$$W_2 = W_{2(-)} + W_{2(+)} = 3 + 4 = 7.$$

На рис. 4 точки W_1, W_2 обведены.

Для ДР (y, x_4) при $Me_{4(-)} = Me_{4(+)}$ с учетом замечания можно принять $W_4 = 2 + 2 = 4$.

В качестве обобщенного критерия Q, объединяющего вклад B и выделившиеся точки W, можно рассматривать

$$Q_i = c\delta B_i + (1-c)\delta W_i,$$

здесь

$$\delta B_i = \frac{|B_i|}{y_{\max} - y_{\min}}; \quad \delta W_i = \frac{W_i}{N}.$$

Весовой коэффициент c может иметь значения [0; 1], рекомендуется брать $c \in [0{,}4;\ 0{,}6]$.

В нашем примере при $c = 0{,}4$

$$Q_1 = 0{,}4\frac{25}{40} + 0{,}6\frac{7}{12} = 0{,}25 + 0{,}35 = 0{,}6;$$

$$Q_2 = 0{,}4\frac{20}{40} + 0{,}6\frac{7}{12} = 0{,}20 + 0{,}35 = 0{,}55;$$

$$Q_3 = 0; \quad Q_4 = 0 + 0{,}6\frac{4}{12} = 0{,}2;$$

$$Q_5 = 0{,}4\frac{5}{40} + 0{,}6\frac{2}{12} = 0{,}05 + 0{,}1 = 0{,}15,$$

т.е. переменная y имеет наибо-лее существенные связи с x_1 и x_2, это показывают значения B_i, W_i и Q_i.

Замечание 2 Если на ДР (y, x_i) для двух уровней значение показателя Q_i мало, но из ана-лиза связей по другим источ-никам переменная x_i должна оказывать сильное влияние на y, то необходимо построить ДР для трех уровней, при этом вводится две промежуточные зоны.

Пример 2 Построим ДР (y, x_5) для трех уровней с двумя промежуточными зонами: первая - $[4{,}5;\ 5{,}5]$, вторая - $[8{,}5;\ 9{,}5]$, значе-ния y, попавшие в эти зоны, исключаются из ДР. Полученная ДР показана на рис. 5. По диаграмме находятся средние значения вклада $|\overline{B_5}|$ и числа выделившихся точек W_5, т.е.

$$|\overline{B_5}| = (40 + |-30|)/2 = 35; \quad \overline{W_5} = \frac{5+6}{2} = 5{,}5.$$

Показатель Q_5 теперь равен:

$$Q_5 = 0{,}4\frac{35}{40} + 0{,}6\frac{5{,}5}{12} = 0{,}35 + 0{,}275 = 0{,}625.$$

Таким образом, новое значение Q_5 превышает Q_1 и Q_2, следовательно, между y, x_5 имеется сильная нелинейная связь.

В общем случае выделение существенных связей между переменными исследуемой системы методом ДР производится в следующей последовательности.

1 Составляется полный перечень входных x_i и выходных переменных y_j системы. Здесь же определяются примерная погрешность Δx_i в регистрации входов.

2 Собирается статистический материал о значениях x_i и y_j за некоторый интервал времени функционирования системы. При этом должны выполняться ряд условий.

Во-первых, размах варьирования переменных x_i, т.е. $x_i^{\max} - x_i^{\min}$, должен в несколько раз превышать погрешность измерения (определения) Δx_i. При этом в случае выбора данных для обработки из большого статистического материала следует отдавать предпочтение значениям x_i наиболее удаленных от их средних $\bar{x}_i$.

Не рекомендуется использовать многократно повторяющиеся наборы значений ($x_{i,j}$, y_j).

Во-вторых, число наблюдений N должно быть таким, чтобы при построении ДР с каждой стороны от ″центральной″ зоны было не менее 4 - 6 точек. Обычно это условие выполняется, если $N > 15 \ldots 20$.

3 Для каждой непрерывно изменяющейся в интервале $\left[x_i^{\max} \ldots x_i^{\min}\right]$ входной переменной x_i рассчитывается оценка математического ожидания по формуле

$$\bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij},$$

x_{ij} - j-е значение входа x_i.

Определяются "центральные" зоны

$$\left[\bar{x}_i - k\Delta x_i; \bar{x}_i + k\Delta x_i\right],\ i = 1, \ldots, n,$$

содержащие значения x_{ij} вблизи оценок $\bar{x}_i$. Коэффициент k = 1 … 3 берется таким, чтобы в центральную зону попадало не более $N/3$ наблюдений.

Если x_i изменяется дискретно, то в мертвую зону включают 1 - 3 дискретных значений центральной части. Так, в примере 1 x_1 может принимать только значения 1, 2, 3, 4, 5. Здесь в "центральную" зону входят x_{ij}=3.

Переменная x_2 изменяется непрерывно, для нее $\bar{x}_2 = 40$ и пусть $\Delta x_2 = 2{,}5$. Тогда в качестве "центральной" зоны можно взять интервал

$$\left[\bar{x}_2 - 2\Delta x_2;\ \ \bar{x}_2 + 2\Delta x_2;\right] = \left[35;\ 45\right].$$

Оценки $\bar{x}_i$ и центральные зоны следует указать в нижней части таблицы исходных данных, а значения x_{ij}, $i = \overline{1,n}$, $j = \overline{1,N}$, попадающие в "центральные" зоны, рекомендуется выделить.

4 Выполняется построение диаграмм рассеяния для всех входных переменных. Если выходных переменных несколько, то ДР строятся для каждой из них.

5 Рассчитываются и на ДР выделяются медианные значения $Me_{i(-)}$, $Me_{i(+)}$, $i = \overline{1,n}$ по формуле (1), определяются величины вкладов входных переменных B_i и числа выделившихся точек W_i по формулам (2), (3), а также значения обобщенного показателя Q_i. Величины B_i и W_i, $i = \overline{1,n}$ записываются под соответствующими ДР.

6 На основании значений B_i, W_i, Q_i делаются выводы: а) какие переменные x_i оказывают наиболее сильное влияние на y; б) каково направление этого влияния; в) какие переменные следует оставить для более детального исследования их связей и построения модели; г) согласуются ли полученные результаты с априорными представлениями.

При необходимости (см. замечание 2) строится ДР для трех уровней по методике, рассмотренной в примере 2.

3 КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

При линейном корреляционном анализе различают следующие виды зависимостей:

1) *парная корреляция* - связь между двумя признаками (y и x, или x_1 и x_2);

2) *частная корреляция* - зависимость между y и одним фактором x_i при фиксированном значении других факторов;

3) *множественная корреляция* - зависимость y и двух или более факторных признаков.

Корреляционный анализ позволяет решать следующие задачи:

1) оценку тесноты связи между показателями с помощью парных, частных и множественных коэффициентов корреляции;

2) оценку уравнения регрессии.

Основная предпосылка применения корреляционного анализа состоит в том, что совокупность значений факторов $(x_1, \ldots, x_k)$ и результативного признака y должна подчиняться нормальному распределению или быть близкой к нему.

3.1 Парная корреляция

Коэффициент парной корреляции r_{xy} между x и y рассчитывается по формуле

$$r_{xy} = \frac{1}{\sigma_x \sigma_y}\left(\frac{1}{N}\sum_{j=1}^{N} x_j y_j - m_x m_y\right) = \frac{K_{x,y}}{\sigma_x \sigma_y}, \qquad (4)$$

$K_{x,y}$ - корреляционный момент (ковариация); m_x, m_y, σ_x, σ_y - средние и средние квадратические отклонения для x и y соответственно, т.е.

$$m_x = \frac{1}{N}\sum_{j=1}^{N} x_j; \quad m_y = \frac{1}{N}\sum_{j=1}^{N} y_j;$$

$$\sigma_x^2 = \frac{1}{N}\sum_{j=1}^{N} x_j^2 - m_x^2; \quad \sigma_y^2 = \frac{1}{N}\sum_{j=1}^{N} y_j^2 - m_y^2 \qquad (5)$$

Из значений коэффициентов парной корреляции составляется корреляционная матрица, например, для y и x_1, x_2, x_3 эта матрица имеет вид

$$R = \begin{array}{c} \\ y \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cccc} y & x_1 & x_2 & x_3 \end{array} \\ \left\| \begin{array}{cccc} 1 & r_{yx_1} & r_{yx_2} & r_{yx_3} \\ r_{x_1y} & 1 & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2y} & r_{x_2x_1} & 1 & r_{x_2x_3} \\ r_{x_3y} & r_{x_3x_1} & r_{x_3x_2} & 1 \end{array} \right\| \end{array}. \qquad (6)$$

Так как коэффициенты r_{yx_i} и r_{x_iy} равны, то матрица может записываться как треугольная, т.е.

$$R = \begin{array}{c} \\ y \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cccc} y & x_1 & x_2 & x_3 \end{array} \\ \left\| \begin{array}{cccc} 1 & r_{yx_1} & r_{yx_2} & r_{yx_3} \\ & 1 & r_{x_1x_2} & r_{x_1x_3} \\ & & 1 & r_{x_2x_3} \\ & & & 1 \end{array} \right\| \end{array}.$$

Рассмотрим иллюстративный пример ($N = 5$) расчета r_{yx_1}, r_{yx_2} и $r_{x_1x_2}$ по формулам (5), (6).

Пример 3 В табл. 2 приведены статистические данные о переменных x_1, x_2, y и результаты расчетов.

Таблица 2

№	y	x_1	x_2		y^2	x_1^2	x_2^2		yx_1	yx_2	x_1x_2
1	5	1	3		25	1	9		5	15	3
2	3	0	5		9	0	25		0	15	0
3	4	2	1		16	4	1		8	4	2
4	4	1	3		16	1	9		4	12	3
5	2	0	0		4	0	0		0	0	0
$\sum_j$	18	4	12	$\sum_j$	70	6	44	$\sum_j$	17	46	8
$m_y\,(m_x)$	3,6	0,8	2,4	σ^2	1,04	0,56	3,04	K_{yx}	0,52	0,56	0,32
				σ	1,02	0,75	1,74	r_{yx}	0,68	0,32	0,25

Примечание

$$m_{x_1} = \frac{1}{5}\sum_{j=1}^{5} x_j = 0{,}8; \qquad m_y = 3{,}6\,; \qquad m_{x_2} = 2{,}4\,;$$

$$\sigma_y^2 = \frac{1}{5}\sum_{j=1}^{5} y_j^2 - m_y^2 = \frac{70}{5} - 3{,}6^2 = 1{,}04\,; \qquad \sigma_y \approx 1{,}02;$$

$$\sigma_{x_1}^2 = 0{,}56; \quad \sigma_{x_1} \approx 0{,}75; \quad \sigma_{x_2}^2 = 3{,}04; \quad \sigma_{x_2} \approx 1{,}74;$$

$$K_{yx_1} = \frac{1}{5}\sum_{i=1}^{5} y_j x_{1j} - m_y m_{x_1} = \frac{17}{5} - 3{,}6 \cdot 0{,}8 = 0{,}52\,;$$

$$r_{yx_1} = \frac{1}{\sigma_y \sigma_{x_1}} K_{yx_1} = \frac{0{,}52}{1{,}02 \cdot 0{,}75} = 0{,}68\,;$$

$$K_{yx_2} = 0{,}56; r_{yx_2} = 0{,}32; K_{x_1x_2} = -0{,}32; r_{x_1x_2} = -0{,}25\,.$$

Корреляционная матрица имеет вид

$$R = \begin{array}{c} \\ y \\ x_1 \\ x_2 \end{array} \begin{array}{c} \begin{array}{ccc} y & x_1 & x_2 \end{array} \\ \left\| \begin{array}{ccc} 1 & 0{,}68 & 0{,}32 \\ & 1 & -0{,}25 \\ & & 1 \end{array} \right\| \end{array}.$$

По данным корреляционной матрицы можно сделать вывод, что между компонентами y и x_1 существует достаточная положительная связь, между y и x_2 - слабая положительная, а между x_1 и x_2 - слабая отрицательная связь.

3.2 Множественная корреляция

Для оценки тесноты связи выходной переменной y с двумя входными переменными x_i и x_j рассчитывается выборочный совокупный коэффициент корреляции $R_{yx_ix_j}$ или R_{yij} по формуле

$$R_{yij} = \sqrt{\frac{r_{yi}^2 + r_{yj}^2 - 2r_{yi}r_{yj}r_{ij}}{1 - r_{ij}^2}}, \qquad (7)$$

$r_{yi} = r_{yx_i}, r_{yj} = r_{yx_j}, r_{ij} = r_{x_ix_j}$.

Величина $R_{yij}^2 = B_{yij}$ называется коэффициентом линейной детерминации, он является мерой линейной связи между y и совокупностью входов x_i, x_j.

В общем случае оценку $B_{yx_ix_jx_k}$ можно произвести с помощью матрицы R по формуле

$$B_{yx_ix_jx_k} = 1 - \frac{\Delta}{\Delta_y},$$

где Δ - определитель матрицы R; Δ_y - определитель матрицы R_y; получающейся из R вычеркиванием первой строки и первого столбца.

При рассмотрении множественной корреляции рекомендуется оценивать тесноту связи между y и x_i при постоянном x_j, т.е. рассчитывать частные коэффициенты корреляции

$$r_{yx_i(x_j)} = \frac{r_{yx_i} - r_{x_ix_j}r_{yx_j}}{\sqrt{\left(1 - r_{x_ix_j}^2\right)\left(1 - r_{yx_j}^2\right)}}. \qquad (8)$$

Пример 3,*а* Для условий примера 3 коэффициенты $B_{yx_1x_2}$ и $R_{yx_1x_2}$ равны

$$R_{yx_1x_2}^2 = B_{yx_1x_2} = \frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2} =$$

$$= \frac{0{,}68^2 + (0{,}32)^2 - 2 \cdot 0{,}68 \cdot 0{,}32(-0{,}25)}{1 - 0{,}25^2} \approx 0{,}7;$$

$$R_{yx_1x_2} = \sqrt{B_{yx_1x_2}} \approx 0{,}84.$$

Расчет частных коэффициентов корреляции $r_{yx_1(x_2)}$ и $r_{yx_2(x_1)}$ показывает

$$r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{x_1x_2}r_{yx_2}}{\sqrt{\left(1 - r_{x_1x_2}^2\right)\left(1 - r_{yx_2}^2\right)}} = \frac{0{,}68 - 0{,}32 \cdot (-0{,}25)}{\sqrt{\left(1 - 0{,}25^2\right)\left(1 - 0{,}32^2\right)}} \approx 0{,}83;$$

$$r_{yx_2(x_1)} = \frac{r_{yx_2} - r_{x_1x_2}r_{yx_1}}{\sqrt{\left(1 - r_{x_1x_2}^2\right)\left(1 - r_{yx_1}^2\right)}} = \frac{0{,}32 - (-0{,}25) \cdot 0{,}68}{\sqrt{\left(1 - 0{,}25^2\right)\left(1 - 0{,}68^2\right)}} \approx 0{,}97.$$

Таким образом, из расчета по данным табл. 2 видно, что между y и x_1, x_2 имеется достаточно сильная линейная связь. Исключение влияния x_1 привело к значительному увеличению связи между y и x_2, так как коэффициент частной корреляции $r_{yx_2(x_1)}$ увеличился по сравнению с коэффициентом парной корреляции r_{yx_2} по абсолютному значению почти в 3 раза.

3.3 Проверка статистических гипотез

Существенность коэффициента множественной корреляции проверяется по F - критерию Фишера, расчетное значение критерия $\hat{F}$ вычисляется по формуле

$$\hat{F} = \frac{B}{1 - B}\frac{N - n_1}{n_1 - 1}, \qquad (9)$$

где n_1 - общее число переменных, например, для $B_{y \cdot x_1x_2}$ $n_1 = 3$.

Если $\hat{F}$ не меньше табличного значения, т.е.

$$\hat{F} \geq F^{\text{т}}_{(\alpha,\, \nu_1 = n_1 - 1,\, \nu_2 = N - n_1)}, \tag{10}$$

то B значимо отличается от нуля и между y и x_1, x_2 имеется существенная линейная связь, здесь α - уровень значимости, ν_1, ν_2 - числа степени свободы.

Значимость частных коэффициентов корреляции можно проверить по *t*-критерию Стьюдента. Для этого вычисляется

$$\hat{t} = \frac{r_{yi(j)}}{\sqrt{1 - r^2_{yi(j)}}} \sqrt{N - n_1} \tag{11}$$

и сравнивается с табличным $t^{\text{т}}_{(\alpha, \nu = N - n_1)}$.

Если $\left|\hat{t}\right| \geq t^{\text{т}}_{\alpha,\nu}$, то связь между y считается существенной.

3.4 Уравнение линейной связи

В случае существенности коэффициента множественной корреляции $R_{yx_1x_2}$ может рассчитываться модель в виде уравнения линейной связи (уравнения регрессии)

$$y = b_0 + b_1 x_1 + b_2 x_2, \tag{12}$$

где b_i - параметры (коэффициенты регрессии).

Параметры b_i можно рассчитать методом наименьших квадратов или по формулам:

$$b_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{\sigma_y}{\sigma_{x_1}};$$

$$b_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{\sigma_y}{\sigma_{x_2}}; \tag{13}$$

$$b_0 = \bar{y} - b_1 m_{x_1} - b_2 m_{x_2}.$$

Таким образом, метод множественной корреляции позволяет:

- оценить тесноту связи между выходной переменной y и входными переменными x_i, $i = 1, 2, \ldots$;
- оценить тесноту связи между y и x_i при постоянном значении x_j;
- рассчитать уравнение связи.

3.5 Нелинейная связь

В случае нелинейной связи например, (см. рис. 3) теснота связи определяется с помощью корреляционного отношения η при рассмотрении двух переменных y и x $(n_1 = 2)$.

Корреляционным отношением y к x, обозначается η_{yx}, называется отношение межгруппового среднего квадратического отклонения $\sigma_{\text{межгр}} = \sigma_{\bar{yx}}$ к общему среднему квадратическому отклонению σ_y.

Корреляционное отношение η_{yx} оценивается по формуле

$$\eta^2_{yx} = \frac{\frac{1}{N}\sum_{j=1}^{k} N_j \left(\bar{y}_j - \bar{y}\right)^2}{\sigma^2_y} = \frac{\sigma^2_{\bar{y}(x)}}{\sigma^2_y}, \tag{14}$$

$\sigma^2_{\bar{y}(x)}$ характеризует рассеяние частных средних $\bar{y}_i$ около общего среднего m_y; N_j - число наблюдений y при неизменном x_j;

$$\bar{y}_j = m_{y_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{ji}, \tag{15}$$

здесь y_{ji} - *i*-е значение *y* при x_j. Величину $\sigma^2_{\bar{y}(x)}$ называют также *межгрупповой дисперсией*.

Значимость корреляционного отношения η проверяется по *t*-критерию

$$\hat{t} = \frac{\eta\sqrt{N-2}}{1-\eta^2}. \qquad (16)$$

Если $\hat{t} < t_{\alpha,\nu}, \quad \nu = N-2,$ (17)

то η - незначим.

Корреляционное отношение обладает следующими свойствами.

1 Корреляционное отношение изменяется в интервале между 0 и 1, т.е.

$$\eta \in [0; 1],$$

причем $\eta = 0$, если между x и y нет корреляционной связи и $\eta = 1$, если зависимость x и y функциональная.

2 Корреляционное отношение является мерой тесноты связи, как для линейной, так и для криволинейной формы связи. При линейной связи η_{yx} теоретически совпадает с r_{yx}.

3 Для криволинейных зависимостей η_{yx} является единственно правильным измерителем тесноты связи.

Всегда выполняется соотношение $\eta_{yx} \geq |r_{yx}|$. Если $\eta_{yx} = |r_{yx}|$, то имеет место линейная корреляционная зависимость.

Для криволинейной зависимости обычно используют модели

$$\bar{y}_x = a + bx + cx^2.$$

Пример 4 Рассчитаем корреляционное отношение η_{yx}. По данным табл. 2 для $x = x_1$ переменная x принимает три значения: $x_1 = 0$, $x_2 = 1$ и $x_3 = 2$. Им соответствуют значения y (2; 3), (4; 5) и 4 со средними $\bar{y}_1 = 2{,}5$; $\bar{y}_2 = 4{,}5$; $\bar{y}_3 = 4$ (см. рис. 6). Общие средняя и дисперсия для y равны $\bar{y} = 3{,}6$, $\sigma_y^2 = 1{,}04$ (табл. 2).

Из таблицы 2

$x = x_j$	0	1	2
y_{ji}	2; 3	4; 5	4
$\bar{y}_j$	2,5	4,5	4
N_j	2	2	1

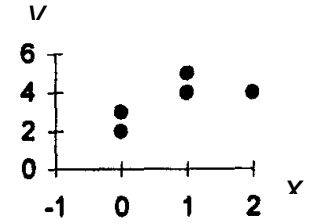


Рис. 6 Корреляционное

Рассчитывая η_{yx}^2 по формуле (14), получаем

$$\eta_{yx}^2 = \frac{1}{\sigma_y^2}\left(\frac{1}{N}\sum_{j=1}^{k} N_j\left(\bar{y}_j - \bar{\bar{y}}\right)^2\right) =$$

$$= \frac{1}{1{,}04}\left[\frac{1}{5}\left(2(2{,}5-3{,}6)^2 + 2(4{,}5-3{,}6)^2 + (4-3{,}6)^2\right)\right] = 0{,}89 \text{ и } \eta_{yx} = 0{,}9.$$

Проверка значимости (существенности) η_{yx} по t -критерию (см. (16) показывает

$$\hat{t} = \frac{0{,}9\sqrt{5-2}}{1-0{,}9^2} = 8{,}2 .$$

Так как

$$\hat{t} = 8{,}2 > t^{\text{т}}_{\alpha=0{,}05;\nu=3} = 3{,}18 ,$$

то корреляционное отношение значимо и между y и x имеет место нелинейная связь.

4 ПРИМЕРЫ ВЫПОЛНЕНИЯ КОНТРОЛЬНОЙ РАБОТЫ (РЕФЕРАТА)

При решении различных задач правовой деятельности с целью определения степени влияния одних переменных на другие для установления тесных связей и выявления наиболее существенных из них используются методы диаграмм рассеяния и корреляционного анализа.

Метод диаграмм рассеяния представляет собой графовычислительный метод, позволяющий без громоздких вычислений определить входные переменные, которые наиболее сильно влияют на изменение выходных переменных.

Метод корреляционного анализа представляет собой аппарат исследования связей между случайными величинами.

4.1 Пример 1 Метод диаграмм рассеяния

1 *Исходные данные.*

За выходную переменную принят процент раскрываемости преступлений в сфере нарушений порядка управления в двенадцати городах, в которых проводились статистические исследования.

Входные переменные:

x_1 - количество подделок документов, государственных наград, штампов, печатей, бланков (на 10 000 человек, с точностью $\Delta x_1 = \pm 0{,}5$);

x_2 - количество похищений или сбыта официальных документов (на 100 000 человек, с точностью $\Delta x_2 = \pm 2{,}5$);

x_3 - количество самоуправств (на 10 000 человек, с точностью $\Delta x_3 = \pm 0{,}1$);

x_4 - количество подделок идентификационного номера транспортного средства (на 10 000 человек, с точностью $\Delta x_4 = \pm 0{,}02$);

x_5 - количество уклонений от прохождения военной и альтернативной службы (на 100 000 человек, с точностью $\Delta x_5 = \pm 0{,}3$).

2 *Входные x_i и выходные y статистические данные приведены в табл. П1.*

3 *По данным табл. 3 строим корреляционные поля (рис. П. 1, а - д).*

Таблица 3

№	x_1	x_2	x_3	x_4	x_5	y
1	5	20	0,32	1,2	7,5	90
2	2	50	0,2	1,1	5	60
3	1	40	0,3	1,15	12	60
4	3	30	0,27	0,9	5	70
5	3	40	0,3	1	9	80
6	4	30	0,15	0,9	7	80
7	1	50	0,4	0,9	3	50
8	4	50	0,5	0,975	12	70
9	3	60	0,4	1,08	11	60
10	2	20	0,1	1	9	70
11	2	60	0,1	0,85	2	50
12	4	30	0,5	1,1	6,5	90

$\bar{x}_i$	3	40	0,3	1	7,5	
Δx_i	0,5	2,5	0,1	0,02	0,3	

$\bar{x}_i$ - среднее значение x_i; Δx_i - ошибка в определении x_i.

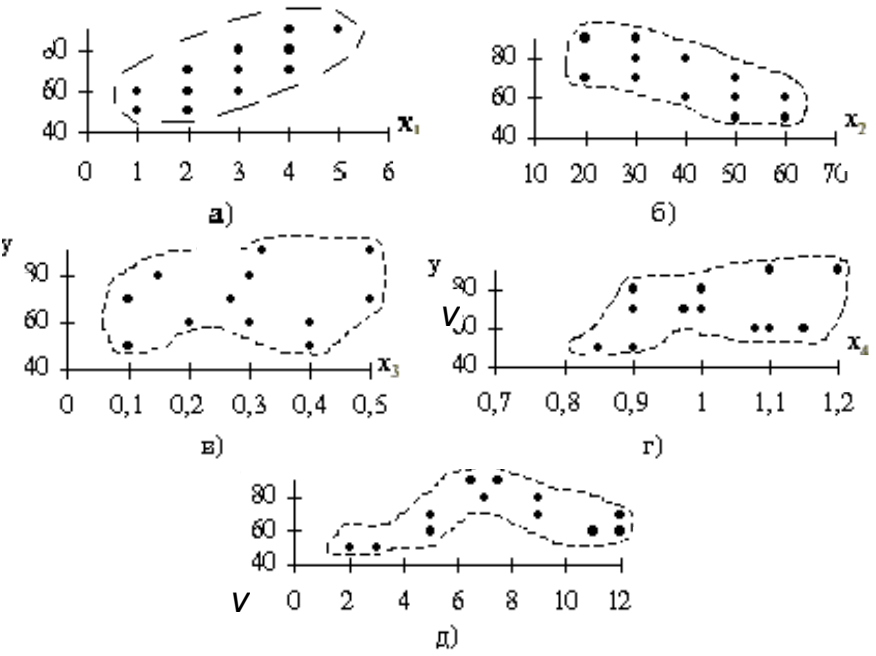


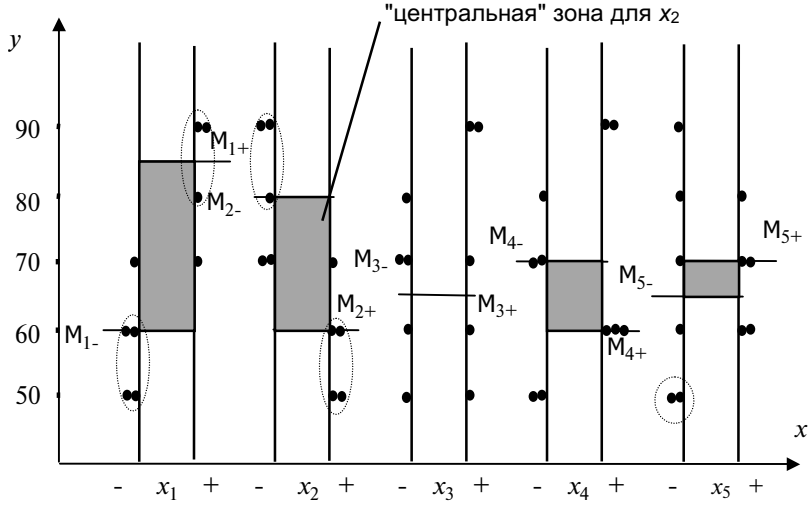
Рис. 7 Корреляционные поля

По виду корреляционного поля можно сделать следующие выводы: на рис. 7, *а* существует линейная, достаточно тесная положительная связь; на рис. 7, *б* - линейная, достаточно тесная отрицательная связь; на рис. 7 *в*, *г* - линейная положительная связь и на рис. 7, *д* - не линейная связь.

4 *Построение диаграмм рассеяния*

Для построения диаграмм рассеяния определяем средние значения $\bar{x}_i$ (см. табл. 3) и значения *y*, соответствующие им, исключаем из дальнейшего рассмотрения (значения x_i, соответствующие среднему значению $\bar{x}_i$ выделены в табл. 5 курсивным (наклонным) шрифтом), оставшиеся значения откладываем на графике (рис. 8) в виде совокупностей точек:

- слева полосы откладываем значения y при $x_i < m_x$;
- справа полосы откладываем значения y при $x_i > m_x$.



$_i$: % 25 - 20 0 -10 5

W_i: 7 7 0 0 2

. 2.

Значения зоны исключения результатов - "центральной" зоны рассчитывается по формуле (П1)

$$[\bar{x}_i - k\Delta x_i; \bar{x}_i + k\Delta x_i], i = 1,\ldots,n, \qquad (П1)$$

содержащие значения x_{ij} вблизи оценок $\bar{x}_i$. Коэффициент k=1÷3 берется таким, чтобы в центральную зону попадало не более $N/3$ наблюдений.

5. Определение вкладов и чисел выделившихся точек.

Существенность влияния входа x_i на выход y определяется двумя параметрами: вкладом и числом выделившихся точек. Вклад рассчитывается по формуле (П2)

$$B_i = Me_{i(+)} - Me_{i(-)}. \qquad (П2)$$

Для определения вклада сначала находим медианные значения $Me_{i(-)}$, $Me_{i(+)}$ для левой и правой совокупности точек ДР. При расчете медианы некоторой совокупности значений y_j, предварительно эти значения записываются в виде ранжированного ряда, т.е. в порядке возрастания (или убывания), количество точек просчитывается и если их число нечетное, то в качестве медианы берется средняя точка, а если число точек четное, то медиана находится между двумя средними точками. Пусть имеется ранжированный ряд $y_1, y_2, \ldots y_n$, тогда

$$Me = \begin{cases} y_{(n+1)/2}, & \text{ñëó÷àéíîå} \\ \frac{1}{2}\left(y_{n/2} + y_{n/2+1}\right), & \text{ñëó÷àéíîå} \end{cases} \qquad (П3)$$

Выделившимися точками слева $W_{i(-)}$ и справа $W_{i(+)}$ для ДР (y, x_i) называются:

1) если левая медиана ниже правой $(Me_{i(-)} < Me_{i(+)})$, то $W_{i(-)}$ образуют точки левой совокупности, находящиеся ниже наименьшего значения точек правой совокупности, а $W_{i(+)}$ образуют точки справа, расположенные выше наибольшего значения точек слева;

2) если левая медиана выше правой $(Me_{i(-)} > Me_{i(+)})$, то $W_{i(-)}$ образуют точки левой совокупности, находящиеся выше наибольшего значения точек справа, а $W_{i(+)}$ - точки правой совокупности, расположенные ниже наименьшего значения точек слева.

Общее число выделившихся точек для ДР (y, x_i) равно

$$W_i = W_{i(-)} + W_{i(+)}. \qquad (П4)$$

Чем больше W_i, тем сильнее влияние x_i на y.

В качестве обобщенного критерия Q, объединяющего вклад B и выделившиеся точки W, можно рассматривать

$$Q_i = c \cdot \delta B_i + (1-c)\delta W_i, \qquad (П5)$$

здесь

$$\delta B_i = \frac{|B_i|}{y_{\max} - y_{\min}}, \delta W_i = \frac{W_i}{N}.$$

Весовой коэффициент с может иметь значения [0; 1], рекомендуется брать $c \in [0{,}4; 0{,}6]$.

В нашем примере при $c = 0{,}4$, $N = 12$.

Расчетные значения $Me_{i(-)}$, $Me_{i(+)}$, B_i, W_i и Q_i даны в Табл. П2.

Таблица П2

	x_1	x_2	x_3	x_4	x_5
$Me_{i(+)}$	85	60	60	60	70
$Me_{i(-)}$	60	80	60	70	65
B_i	25	-20	0	-10	5
W_i	7	7	0	0	2
Q_i	0,6	0,55	0	0,1	0,15

$$Q_1 = 0{,}4 \cdot \frac{25}{90-50} + 0{,}6 \cdot \frac{7}{12} = 0{,}25 + 0{,}35 = 0{,}6 \quad \text{и т.д.}$$

Вывод:

1) исходя из величин вкладов B_i и чисел выделившихся точек W_i, определились две входные переменных x_1 и x_2, им соответствуют B_1=25, W_1=7 и B_2=-20, W_2=7, которые имеют существенное влияние на выходную переменную y ;

2) с увеличением x_1 - y увеличивается, с увеличением x_2 - y уменьшается.

Учебное издание

ИНФОРМАТИКА

(*корреляционный анализ и метод диаграмм рассеяния*)

Методические указания

Авторы-составители: **МУРОМЦЕВ** Юрий Леонидович
ОРЛОВА Лариса Павловна
БУРЦЕВА Елена Васильевна
МУРОМЦЕВ Дмитрий Юрьевич

Редактор Т. М. Глинкина
Инженер по компьютерному макетированию
Г. Ю. Корабельникова

ИНФОРМАТИКА

(*корреляционный анализ и*
метод диаграмм рассеяния)

Издательство ТГТУ