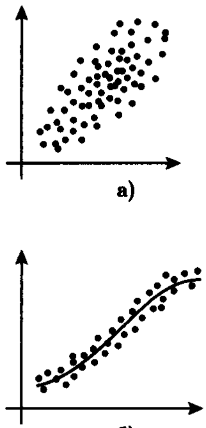
ГЛАВА 20

КОРРЕЛЯЦИЯ

Сократ. А смог бы ты, не глядя на скалу, а рассматривая ее отражение в воде, сказать, как можно было бы взобраться на самую вершину?

А. Реньи. Диалоги

После того, как (с помощью методов из гл. 19) объекты разбиты на однородные группы (классы), возникает задача изучения *взаимосвязей признаков* внутри отдельного класса. На практике чаще всего встречаются следующие два вида зависимостей: а) объекты образуют «облако» эллиптического типа (рис. 1, *а*), б) объекты располагаются в окрестности некоторой кривой (поверхности) (рис. 1, *б*). В случае а) оба признака являются «полноценными» случайными величинами, и изучению подлежит уровень зависимости (корреляции) между ними. Случай б) соответствует «функциональной» зависимости между признаками, испорченной шумом. Зависимости первого вида изучаются в этой главе методами *корреляционного анализа*. Методы, позволяющие во втором случае построить интересующую исследователя кривую (поверхность), относятся к так называемому *регрессионному анализу*. Они обсуждаются в гл. 21 и § 2 гл. 22.



а) б)

Рис. 1

§ 1. ГЕОМЕТРИЯ ГЛАВНЫХ КОМПОНЕНТ

Допустим, что каждый из n объектов описывается m признаками (координатами), и представим данные (для отдельного класса объектов) в форме таблицы $\boldsymbol{X} = \|x_{il}\|_{n\times m}$. Вычислим для каждого признака (столбца матрицы $\boldsymbol{X}$) *среднее значение* $\overline{x}_l = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_{il}$ и центрируем данные: $x'_{il} = x_{il} - \overline{x}_l$.*) Далее в этой главе будем считать x_{il} уже *центрированными*:

Д1. $\overline{x}_l = 0$ для $l = 1,\ldots,m$.

Обозначим через $\widehat{\boldsymbol{\Sigma}} = \|\widehat{\sigma}_{kl}\|_{m\times m}$ *выборочную ковариационную матрицу* (центрированных) признаков: $\widehat{\boldsymbol{\Sigma}} = \dfrac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ (т. е. $\widehat{\sigma}_{kl}$ — выборочная ковариация k-го и l-го столбцов матрицы $\boldsymbol{X}$).

Поскольку $\widehat{\boldsymbol{\Sigma}}$ — матрица ковариаций, она неотрицательно определена (см. П10). Следовательно, существует ортогональная матрица $\boldsymbol{C}$, приводящая $\widehat{\boldsymbol{\Sigma}}$ к главным осям: $\boldsymbol{C}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{C} = \boldsymbol{\Lambda}$. Здесь $\boldsymbol{\Lambda}$ — диагональная матрица с неотрицательными элементами $\lambda_1 \geqslant \ldots \geqslant \lambda_m$ на главной диагонали, которые являются корнями уравнения $\det(\widehat{\boldsymbol{\Sigma}} - \lambda\boldsymbol{E}) = 0$. Они называются *собственными значениями* матрицы $\widehat{\boldsymbol{\Sigma}}$. Предположим, что все λ_l *положительны*

*) Это преобразование не искажает интересующую нас внутреннюю структуру класса, характер взаимосвязей признаков.

и различны (для экспериментальных данных x_{il} это условие выполняется практически всегда). При этом столбцы $\boldsymbol{c}_1,\dots,\boldsymbol{c}_m$ матрицы $\boldsymbol{C}$ (*главные оси или компоненты*) определяются однозначно с точностью до выбора направления оси (одновременного изменения знака всех координат вектора $\boldsymbol{c}_l$). Они образуют новый ортонормированный базис в $\mathbb{R}^m$ (рис. 2), обладающий **рядом важных свойств**.

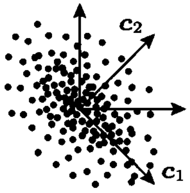


Рис. 2

1) Проекции объектов на первую главную компоненту $\boldsymbol{c}_1$ имеют *наибольшую* выборочную дисперсию среди проекций на всевозможные направления $\boldsymbol{d}$ в пространстве $\mathbb{R}^m$.

Доказательство. Вектор проекций $\boldsymbol{y}$ на направление $\boldsymbol{d}$ ($\boldsymbol{d}^T\boldsymbol{d}=1$) задается равенством $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{d}$. Ввиду допущения Д1

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m}x_{il}d_l=\sum_{l=1}^{m}\overline{x}_l d_l=0.$$

Тогда *выборочная дисперсия проекций на направление* $\boldsymbol{d}$ равна

$$S^2(\boldsymbol{d})=\frac{1}{n}\sum_{i=1}^{n}y_i^2=\frac{1}{n}\,\boldsymbol{y}^T\boldsymbol{y}=\frac{1}{n}\,(\boldsymbol{X}\boldsymbol{d})^T\boldsymbol{X}\boldsymbol{d}=\boldsymbol{d}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{d}. \qquad (1)$$

Тем самым задача сводится к максимизации по $\boldsymbol{d}$ квадратичной формы $\boldsymbol{d}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{d}$ при условии $\boldsymbol{d}^T\boldsymbol{d}=1$. Для ее решения применим метод неопределенных множителей Лагранжа (см. [46, с. 271]). Приравнивая нулю частные производные по переменным d_l *функции Лагранжа*

$$F(\boldsymbol{d},\lambda)=\boldsymbol{d}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{d}-\lambda(\boldsymbol{d}^T\boldsymbol{d}-1)=\sum_{k=1}^{m}\sum_{l=1}^{m}\widehat{\sigma}_{kl}d_k d_l-\lambda\sum_{l=1}^{m}d_l^2+\lambda,$$

приходим к системе (линейных относительно $d_1,\dots,d_m$) уравнений

$$\frac{\partial}{\partial d_l}\,F(\boldsymbol{d},\lambda)=2\sum_{k=1}^{m}\widehat{\sigma}_{kl}d_k-2\lambda d_l=0,\quad l=1,\dots,m.$$

Ее можно записать в матричной форме:

$$\widehat{\boldsymbol{\Sigma}}\boldsymbol{d}-\lambda\boldsymbol{d}=\boldsymbol{0}\iff(\widehat{\boldsymbol{\Sigma}}-\lambda\boldsymbol{E})\boldsymbol{d}=\boldsymbol{0}. \qquad (2)$$

Поскольку $\boldsymbol{d}^T\boldsymbol{d}=1$, нас интересуют только ненулевые решения. Для них должно выполняться условие $\det(\widehat{\boldsymbol{\Sigma}}-\lambda\boldsymbol{E})=0$, т. е. искомое направление $\boldsymbol{d}$ обязано быть собственным вектором $\widehat{\boldsymbol{\Sigma}}$, отвечающим собственному значению λ. Умножая соотношение (2) на $\boldsymbol{d}^T$ слева, получим

$$\boldsymbol{d}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{d}=\lambda\,\boldsymbol{d}^T\boldsymbol{d}=\lambda.$$

Левая часть есть $S^2(\boldsymbol{d})$ (см. формулу (1)). Следовательно, λ должно равняться наибольшему собственному значению λ_1 матрицы $\widehat{\boldsymbol{\Sigma}}$, а $\boldsymbol{d}=\boldsymbol{c}_1$. ■

2) Второй собственный вектор $\boldsymbol{c}_2$ характеризуется тем, что выборочная дисперсия проекций объектов на ось $\boldsymbol{c}_2$ максимальна среди всех направлений $\boldsymbol{d}$, *ортогональных вектору* $\boldsymbol{c}_1$, т. е. таких, что $\boldsymbol{d}^T\boldsymbol{c}_1 = 0$.

   ДОКАЗАТЕЛЬСТВО. Функция Лагранжа в этом случае выглядит так:

   $$F(\boldsymbol{d},\lambda,\mu) = \boldsymbol{d}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{d} - \lambda(\boldsymbol{d}^T\boldsymbol{d} - 1) - \mu(\boldsymbol{d}^T\boldsymbol{c}_1 - 0).$$

   Вычисляя ее частные производные по d_l, приходим к системе

   $$\widehat{\boldsymbol{\Sigma}}\boldsymbol{d} - \lambda\boldsymbol{d} - \mu\boldsymbol{c}_1 = \boldsymbol{0}. \qquad (3)$$

   Транспонируем равенство (3), умножим на $\boldsymbol{c}_1$ справа и воспользуемся тем, что $\boldsymbol{c}_1$ — собственный вектор с собственным значением λ_1 (т. е. $\widehat{\boldsymbol{\Sigma}}\boldsymbol{c}_1 = \lambda_1\boldsymbol{c}_1$), а также условием $\boldsymbol{d}^T\boldsymbol{c}_1 = 0$:

   $$\boldsymbol{d}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{c}_1 - \lambda\,\boldsymbol{d}^T\boldsymbol{c}_1 - \mu\,\boldsymbol{c}_1^T\boldsymbol{c}_1 = 0 \iff \lambda_1\cdot 0 - \lambda\cdot 0 - \mu\cdot 1 = 0.$$

   Отсюда $\mu = 0$. С учетом этого из (3) вытекает, что $\boldsymbol{d}$ — собственный вектор матрицы $\widehat{\boldsymbol{\Sigma}}$ с собственным значением λ. Умножая равенство (3) слева на $\boldsymbol{d}^T$, выводим, что $\lambda = \lambda_2$ и $\boldsymbol{d} = \boldsymbol{c}_2$. ■

3) При $l \geqslant 3$ аналогично устанавливается, что $\boldsymbol{c}_l$ — направление с наибольшей выборочной дисперсией проекций объектов среди направлений, ортогональных векторам $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{l-1}$.

4) Сумма выборочных дисперсий исходных признаков (столбцов матрицы $\boldsymbol{X}$) $\widehat{\sigma}_{11} + \ldots + \widehat{\sigma}_{mm} = \operatorname{tr}\widehat{\boldsymbol{\Sigma}}$ в силу подобия матриц $\widehat{\boldsymbol{\Sigma}}$ и $\boldsymbol{\Lambda}$ (см. П10) равна $\operatorname{tr}\boldsymbol{\Lambda} = \lambda_1 + \ldots + \lambda_m = S^2(\boldsymbol{c}_1) + \ldots + S^2(\boldsymbol{c}_m)$, т. е. сумме выборочных дисперсий проекций объектов на (новые) главные оси. Эта величина может рассматриваться как *мера общего разброса* объектов относительно их центра масс. Представляет интерес *относительная доля разброса, приходящаяся на k первых главных осей*,

   $$\gamma_k = (\lambda_1 + \ldots + \lambda_k)/(\lambda_1 + \ldots + \lambda_m), \quad k \leqslant m.$$

   Если эта величина при некотором k достаточно близка к 1, то возможно *уменьшение размерности* пространства признаков за счет перехода от m исходных признаков к k новым признакам — первым главным компонентам. На практике нередко удается ограничиться двумя или тремя компонентами без существенной потери информации. Объекты описываются координатами в новых осях, которым специалисты-прикладники, как правило, могут придать содержательную интерпретацию.

Математика приводит нас к дверям истины, но самих дверей не открывает.

В. Ф. Одоевский

Пример 1 ([30, с. 206]). Найдем и интерпретируем главные компоненты для данных примера 4 гл. 19. Напомним, что исходными признаками там были следующие: 1 — длина черепа, 2 — длина верхней челюсти, 3 — ширина верхней челюсти, 4 —

	1	2	3	4	5	6
1	1	0,96	0,35	0,61	0,72	0,59
2		1	0,20	0,66	0,74	0,59
3			1	0,37	0,35	0,35
4				1	0,89	0,76
5					1	0,79
6						1

Рис. 3

c_1	c_2	c_3	c_4	c_5	c_6
0,43	0,23	0,53	0,11	0,05	0,68
0,43	0,38	0,39	0,01	−0,20	−0,69
0,23	−0,89	0,38	−0,02	0,00	−0,13
0,44	−0,07	−0,40	−0,52	−0,58	0,18
0,46	0,02	−0,27	−0,31	0,78	−0,09
0,42	−0,10	−0,44	0,79	−0,09	−0,01

Рис. 4

длина верхнего карнивора, 5 — длина первого верхнего моляра, 6 — ширина первого верхнего моляра.

Очевидно, что при нормировке данных с помощью средних арифметических и стандартных отклонений признаков (N2 из § 1 гл. 19) выборочная ковариационная матрица $\widehat{\Sigma}$ совпадает с выборочной корреляционной матрицей исходных признаков. [Ее нетрудно подсчитать с помощью программы Excel. Результаты вычислений представлены таблицей на рис. 3.]

Обратим внимание, что длина черепа (признак 1) и длина верхней челюсти (признак 2) сильно коррелируют ($\widehat{\rho}_{12} = 0{,}96$), поэтому целесообразно оставить в модели только один из них. Признаки 4, 5 и 6, относящиеся к зубам, также очень тесно связаны между собой, поскольку $\widehat{\rho}_{45} = 0{,}89$, $\widehat{\rho}_{46} = 0{,}76$ и $\widehat{\rho}_{56} = 0{,}79$.

На рис. 4 приведены собственные векторы $c_1, \ldots, c_6$, вычисленные *степенным методом* (см. § 3). Собственные значения λ_k, соответствующие им доли следа $\lambda_k / \sum \lambda_l$ (в %) и накопленные доли γ_k (в %) указаны в следующей таблице:

Номера компонент	1	2	3	4	5	6
Собственные значения	4,100	0,883	0,639	0,259	0,097	0,022
Проценты от следа	68,3	14,7	10,7	4,3	1,6	0,4
Накопленные проценты	68,3	83,0	93,7	98,0	99,6	100,0

На первые 3 компоненты приходится 93,7% полной дисперсии «облака». При этом первая компонента имеет смысл *общего размера*. Это следует из того, что все координаты у c_1 одного знака и примерно одинаковы по величине, т. е. при проецировании на эту ось координаты нормированных признаков просто складываются.

Вторая компонента, по существу, отвечает за *ширину верхней челюсти* (признак 3), поскольку третья координата у c_2 по абсолютной величине равна 0,89 $\approx$ 1. Эта ось отражает различие в пропорциях челюстей и отличает удлиненные формы от укороченных (гончих и колли от бульдогов и боксеров). На вторую ось волки проецируются в основном рядом с немецкими овчарками.

Третья ось противопоставляет размеры челюстей размерам зубов: первые три координаты у c_3 примерно равны по абсолютной величине последним трем координатам, но противоположны по знаку. Другими словами, третья ось отражает *относительную* (по сравнению с размерами черепа) *величину зубов*. Она позволяет отличить животных с развитыми зубами (волки, немецкие овчарки, доберманы) от собак других пород (сенбернары, мастифы, сеттеры).

5) Пусть M_k — это подпространство, натянутое на главные оси $c_1, \ldots, c_k$. Оказывается, при проецировании объектов на произвольное подпространство L_k размерности k в $\mathbb{R}^m$ геометрическая структура *искажается в наименьшей степени*, если этим подпространством является M_k (см. [1, с. 350]):

а) сумма квадратов расстояний от объектов до их проекций на L_k минимальна, когда $L_k = M_k$ (в этом случае она равна $n(\lambda_{k+1} + \ldots + \lambda_m)$ (доказательство см. в [76, с. 243]));

б) при проецировании на M_k наименее искажается сумма квадратов расстояний между всевозможными парами объектов (для M_k ее изменение составляет $n^2(\lambda_{k+1} + \ldots + \lambda_m)$);

в) когда $L_k = M_k$, в наименьшей степени искажаются расстояния от объектов до их центра масс (совпадающего с началом координат $\mathbf{0}$ ввиду допущения Д1), а также углы между всевозможными парами прямых, соединяющих объекты с $\mathbf{0}$.

Поясним последнее свойство. Рассмотрим *матрицу скалярных произведений* $\boldsymbol{G} = \|g_{ij}\|_{n\times n} = \boldsymbol{X}\boldsymbol{X}^T$. Нетрудно понять геометрический смысл элементов этой матрицы: $g_{ii} = \sum\limits_{l=1}^{m} x_{il}^2$ представляет собой квадрат расстояния от i-го объекта до $\mathbf{0}$, а при $i \neq j$ величина $g_{ij} = \sum\limits_{l=1}^{m} x_{il}x_{jl}$ пропорциональна косинусу угла между прямыми, соединяющими i-й и j-й объекты с началом координат.

Обозначим через $\boldsymbol{Y} = \|y_{il}\|_{n\times m}$ матрицу координат проекций объектов на подпространство L_k. Ей соответствует $\boldsymbol{H} = \|h_{ij}\|_{n\times n} = \boldsymbol{Y}\boldsymbol{Y}^T$. Тогда при $L_k = M_k$ достигается

$$\min_{L_k} |\boldsymbol{G} - \boldsymbol{H}|^2 = n^2(\lambda_{k+1}^2 + \ldots + \lambda_m^2), \tag{4}$$

где $|A|^2 = \sum a_{ij}^2$ — квадрат *евклидовой нормы матрицы* $\boldsymbol{A}$.

Замечание 1. Проецирование на плоскость двух первых компонент часто применяется еще на этапе классификации (выделении однородных групп). Авторы [4, с. 103] считают: «Гипотеза состоит в том, что наибольший разброс данные будут иметь в направлениях, «соединяющих» центры групп, а значит, проекции на старшие главные компоненты обеспечат наилучшую «точку зрения» на данные, когда группы видны на наибольших расстояниях и не закрывают одна другую».

Замечание 2. Следует иметь в виду, что главные компоненты вычисляются по выборочной ковариационной матрице $\widehat{\boldsymbol{\Sigma}}$ и поэтому *зависят от масштаба признаков*. Скажем, если один из признаков принимает значения от 0 до 100, а другие — от 0 до 10, то независимо от структуры данных первый признак будет отождествляться с первой главной компонентой. Чтобы избежать этого, обычно перед вычислением главных компонент данные нормируют (см. § 1 гл. 19).

§ 2. ЭЛЛИПСОИД РАССЕЯНИЯ

Рассмотрим m-мерный случайный вектор $\boldsymbol{\xi}$ с математическим ожиданием $\mathsf{M}\boldsymbol{\xi} = \mathbf{0}$ и ковариационной матрицей $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$. (В частности, годится *эмпирическое распределение**) с нулевыми средними арифметическими значений координат признаков (допущение Д1 из § 1) и выборочной ковариационной матрицей $\widehat{\boldsymbol{\Sigma}}$.)

В П10 доказано, что любая ковариационная матрица неотрицательно определена. Потребуем дополнительно, чтобы матрица $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ была *невырожденной*. Это равносильно ее положительной определенности, а также положительности всех ее собственных значений $\lambda_1 \geqslant \ldots \geqslant \lambda_m$. Обозначим соответствующие им собственные векторы (направления главных осей) через $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m$.

Для произвольного направления $\boldsymbol{d}$ в $\mathbb{R}^m$ ($|\boldsymbol{d}| = 1$) случайная величина $\zeta = \boldsymbol{d}^T\boldsymbol{\xi}$ представляет собой *координату проекции* вектора $\boldsymbol{\xi}$ на направление $\boldsymbol{d}$. Найдем такое направление, для которого дисперсия $\mathsf{D}\zeta$ имеет *наибольшее значение*. Запишем:

$$\mathsf{D}\zeta = \mathsf{M}(\boldsymbol{d}^T\boldsymbol{\xi})^2 = \mathsf{M}(\boldsymbol{d}^T\boldsymbol{\xi})(\boldsymbol{d}^T\boldsymbol{\xi})^T = \mathsf{M}\,\boldsymbol{d}^T\boldsymbol{\xi}\boldsymbol{\xi}^T\boldsymbol{d} = \boldsymbol{d}^T\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{d}.$$

В точности так же, как в § 1 для случая выборочной ковариационной матрицы $\widehat{\boldsymbol{\Sigma}}$, доказывается, что максимум дисперсии $\mathsf{D}\zeta$ достигается на направлении $\boldsymbol{c}_1$.

Аналогично устанавливается, что при $l > 1$ собственный вектор $\boldsymbol{c}_l$ является направлением с наибольшей дисперсией $\mathsf{D}\zeta$ среди направлений, ортогональных векторам $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{l-1}$.

В силу того, что $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ невырождена, существует обратная к ней матрица $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1}$, которая также положительно определена (см. вопрос 6 гл. 19).

Определение. *Эллипсоидом рассеяния* распределения вектора $\boldsymbol{\xi}$ называется m-мерный эллипсоид

$$\boldsymbol{x}^T\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1}\boldsymbol{x} \leqslant m + 2.$$

*) У которого каждому набору $(x_{i1}, \ldots, x_{im})$ координат i-го объекта (т. е. строке таблицы данных $\boldsymbol{X}$) приписана вероятность $1/n$.

Он однозначно выделяется среди всех других эллипсоидов следующим своим свойством (см. [44, с. 333]): если рассмотреть *равномерно распределенный* на нем вектор $\boldsymbol{U}$ (имеющий постоянную плотность внутри эллипсоида и равную 0 вне его), то первые (равные 0) и вторые моменты (т. е. ковариации компонент) вектора $\boldsymbol{U}$ совпадут с моментами вектора $\boldsymbol{\xi}$.

Осями эллипсоида рассеяния служат главные оси матрицы $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ (рис. 5 для $\widehat{\boldsymbol{\Sigma}}$). Длины его полуосей пропорциональны $\sqrt{\lambda_l}$, где λ_l — собственные значения матрицы $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$, а m-мерный объем равен константе, умноженной на $(\det \boldsymbol{\Sigma}_{\boldsymbol{\xi}})^{1/2} = (\lambda_1 \ldots \lambda_m)^{1/2}$.

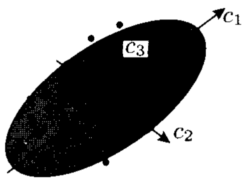


Рис. 5

В заключение параграфа познакомимся с одним из способов сравнения «степеней рассеяния» многомерных распределений, который связан с их эллипсоидами рассеяния.

Пусть m-мерные случайные векторы $\boldsymbol{\xi}$ и $\boldsymbol{\eta}$ имеют распределения с $\mathsf{M}\boldsymbol{\xi} = \mathsf{M}\boldsymbol{\eta} = \mathbf{0}$ и матрицами ковариаций $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$.

Определение. Будем говорить, что *среднеквадратическое рассеяние* случайного вектора $\boldsymbol{\xi}$ вокруг начала координат $\mathbf{0}$ не больше, чем рассеяние вектора $\boldsymbol{\eta}$, если для любого $\boldsymbol{d} \in \mathbb{R}^m$ верно неравенство

$$\mathsf{M}(\boldsymbol{d}^T\boldsymbol{\xi})^2 \leqslant \mathsf{M}(\boldsymbol{d}^T\boldsymbol{\eta})^2. \tag{5}$$

Неравенство (5) означает, что дисперсия случайной величины $\boldsymbol{d}^T\boldsymbol{\xi}$ для любого направления $\boldsymbol{d}$ не превосходит дисперсии случайной величины $\boldsymbol{d}^T\boldsymbol{\eta}$. Раскрывая скобки, очевидно, получаем, что неравенство (5) равносильно неотрицательной определенности матрицы $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}}$.

В [11, с. 102] доказано, что при условии невырожденности матриц $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ и $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ среднеквадратическое рассеяние вектора $\boldsymbol{\xi}$ вокруг начала координат не больше рассеяния вектора $\boldsymbol{\eta}$ тогда и только тогда, когда эллипсоид рассеяния вектора $\boldsymbol{\xi}$ целиком *лежит внутри* эллипсоида рассеяния вектора $\boldsymbol{\eta}$.

Замечание 3. При $m \geqslant 2$ неравенство (5) устанавливает лишь частичный порядок на множестве ковариационных матриц. Например, матрицы $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ и $\begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$ не лучше и не хуже одна другой, поскольку в направлении $\boldsymbol{d} = (1{,}0)$ меньше дисперсия для первой матрицы, а в направлении $\boldsymbol{d} = (0{,}1)$ — для второй (рис. 6).

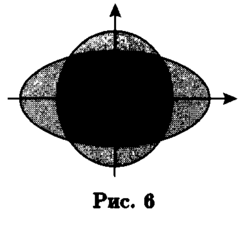


Рис. 6

В подобной ситуации для сравнения «степеней рассеяния» многомерных законов обычно используют такие скалярные характеристики ковариационных матриц, как след $\lambda_1 + \ldots + \lambda_m$ или определитель $\lambda_1 \ldots \lambda_m$. При этом надо иметь в виду, что выводы для разных характеристик (как и для выше приведенных матриц) могут оказаться *прямо противоположными* (эту тему продолжает задача 1).

§ 7. РАНГОВАЯ КОРРЕЛЯЦИЯ

Нередко на практике представляет интерес *гипотеза независимости признаков* ξ и η:

$H_0\colon F_{\xi,\eta}(x,y) = F_\xi(x)F_\eta(y)$ при всех x,y,

где $F_{\xi,\eta}(x,y) = \mathsf{P}(\xi \leqslant x, \eta \leqslant y)$ — это функция распределения случайного вектора (ξ,η) (см. П8), а $F_\xi(x)$ и $F_\eta(y)$ — функции распределения его компонент.

Для ее проверки применяют ранговые критерии. Они не зависят от конкретного вида функций F_ξ и F_η при условии их непрерывности. Кроме того, эти критерии робастны (устойчивы) (см. § 4 гл. 8) к выделяющимся наблюдениям («выбросам»), которые обычно присутствуют в крупных массивах реальных данных. Рассмотрим сначала наиболее часто используемый

Критерий Спирмена

Обозначим через R_i *ранг* (т. е. номер в порядке возрастания) наблюдения ξ_i среди $\xi_1,\ldots,\xi_n$, а через S_i ранг η_i среди $\eta_1,\ldots,\eta_n$. Таким образом, наблюдения порождают n пар рангов $(R_1,S_1),\ldots,(R_n,S_n)$. Статистикой критерия Спирмена служит *выборочный коэффициент корреляции* ρ_S ранговых наборов $(R_1,\ldots,R_n)$ и $(S_1,\ldots,S_n)$, определяемый формулой

$$\rho_S = \sum_{i=1}^n (R_i - \overline{R})(S_i - \overline{S}) \Big/ \left[\sum_{i=1}^n (R_i - \overline{R})^2 \sum_{i=1}^n (S_i - \overline{S})^2\right]^{1/2}. \qquad (20)$$

В этой формуле $\overline{R} = \overline{S} = \frac{1}{n}\sum_{i=1}^n i = \frac{n+1}{2}$. С учетом легко доказываемого по индукции равенства $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$ имеем

$$\sum_{i=1}^n (R_i - \overline{R})^2 = \sum_{i=1}^n (S_i - \overline{S})^2 = \sum_{i=1}^n \left(i - \frac{n+1}{2}\right)^2 = \frac{n^3-n}{12}.$$

Переставив пары (R_i,S_i) в порядке возрастания первой компоненты, получим набор $(1,T_1),\ldots,(n,T_n)$. Тогда статистика (20) запишется в виде

$$\rho_S = \frac{12}{n^3-n}\sum_{i=1}^n \left(i - \frac{n+1}{2}\right)\left(T_i - \frac{n+1}{2}\right). \qquad (21)$$

Таким образом, ρ_S — линейная функция от рангов T_i. Правую часть равенства (21) можно также представить (задача 3) в виде

$$\rho_S = 1 - \frac{6}{n^3-n}\sum_{i=1}^n (i - T_i)^2 = 1 - \frac{6}{n(n-1)(n+1)}\sum_{i=1}^n (R_i - S_i)^2, \qquad (22)$$

который наиболее удобен для вычислений.

Совпадения. Пример 9 показывает, что формула (22) пригодна для подсчета ρ_S *только в случае отсутствия совпадений* (т. е. в случае, когда среди значений наблюдений $\xi_1,\ldots,\xi_n$ $(\eta_1,\ldots,\eta_n)$ нет одинаковых). Если совпадения есть, то при ранжировании им следует присваивать *средние ранги**) и затем вычислять ρ_S на основе формулы (20).**)

Пример 9 ([2, с. 108]). Десять однородных предприятий были проранжированы вначале по *степени прогрессивности их оргструктур* (признак ξ), затем по *эффективности их функционирования в отчетном году* (признак η). В результате были получены следующие две ранжировки: 1; 2,5; 2,5; 4,5; 4,5; 6,5; 6,5; 8; 9,5; 9,5 и 1; 2; 4,5; 4,5; 4,5; 4,5; 8; 8; 8; 10. Для этих данных правая часть равенства (22) равна 0,921, а коэффициент ρ_S, вычисленный по формуле (20), имеет значение 0,917. С помощью табл. Т6 устанавливаем, что при $n = 10$ критической границей для ρ_S на уровне значимости 0,001 служит величина 0,879. Поскольку $0{,}917 > 0{,}879$, корреляционную связь между признаками ξ и η следует признать значимой.

Исследуем некоторые свойства статистики ρ_S при справедливости гипотезы H_0. Множество рангов $(T_1,\ldots,T_n)$ — это некоторая перестановка множества $(1,\ldots,n)$. При выполнении гипотезы H_0 все $n!$ таких перестановок равновероятны. Поэтому для любого $1 \leqslant i \leqslant n$

$$\mathbf{M}T_i = \sum_{k=1}^{n} k\,\mathbf{P}(T_i = k) = \sum_{k=1}^{n} k\,\frac{(n-1)!}{n!} = \frac{n+1}{2}\,.$$

Из формулы (21) немедленно получаем, что $\mathbf{M}\rho_S = 0$ при выполнении гипотезы H_0. Нетрудно установить, что $\mathbf{D}\rho_S = 1/(n-1)$ (задача 4). Так как ρ_S — коэффициент корреляции, то согласно следствию из неравенства Коши—Буняковского (П4) всегда $-1 \leqslant \rho_S \leqslant 1$. Крайние значения достигаются: при полном соответствии наборов рангов ($R_i = S_i$, $i = 1,\ldots,n$) имеем $\rho_S = 1$, а при противоположных рангах ($T_i = n - i + 1$) получаем $\rho_S = -1$.

Вопрос 2.
Почему верно последнее утверждение?

Достаточно близкие к 1 (или -1) значения ρ_S противоречат гипотезе H_0. Критические границы при односторонней альтернативе $H_1\colon \boldsymbol{\rho}(\xi,\eta) > 0$ на нескольких уровнях значимости для $n \leqslant 50$ можно найти в табл. Т6.

Для $n > 50$ годится нормальное приближение, основанное на сходимости

$$\rho_S \Big/ \sqrt{\mathbf{D}\rho_S} \xrightarrow{d} Z \sim \mathcal{N}(0,1) \quad \text{при } n \to \infty$$

в случае справедливости гипотезы H_0 (доказательство см. в [86, с. 227]).

*) Так, для выборки 2, 5, 5, 7 получаем ранжировку 1; 2,5; 2,5; 4.

**) Для подсчета выборочного коэффициента корреляции ранжировок (20) можно воспользоваться функцией «Коррел» из Excel.

Поправка. Для небольших выборок это приближение не является удовлетворительным. Р. Иман и У. Коновер в 1978 г. предложили следующую поправку, значительно повышающую точность аппроксимации (см. [88, с. 10]). Положим

$$\tilde{\rho}_S = \frac{1}{2}\,\rho_S\left(\sqrt{n-1} + \sqrt{(n-2)/(1-\rho_S^2)}\right).$$

С помощью табл. Т2 и Т4 вычислим $z_\alpha = (x_{1-\alpha} + y_{1-\alpha})/2$, где $x_{1-\alpha}$ и $y_{1-\alpha}$ обозначают, соответственно, квантили уровня $(1-\alpha)$ закона $\mathcal{N}(0,1)$ и распределения Стьюдента с $(n-2)$ степенями свободы (см. § 2 гл. 11). Если $\tilde{\rho}_S \geqslant z_\alpha$, то гипотеза H_0 отвергается в пользу альтернативы $H_1\colon \boldsymbol{\rho}(\xi,\eta) > 0$, иначе — принимается.

Критерий Кендэла

Другую ранговую меру связи ввел в 1938 г. М. Дж. Кендэл. Будем говорить, что пары (ξ_i,η_i) и (ξ_j,η_j) *согласованы* $(1 \leqslant i < j \leqslant n)$, если $\xi_i < \xi_j$ и $\eta_i < \eta_j$ или $\xi_i > \xi_j$ и $\eta_i > \eta_j$ (т. е. $\operatorname{sign}(\xi_j-\xi_i)\operatorname{sign}(\eta_j-\eta_i) = 1$). Пусть S — число согласованных пар, а R — число несогласованных пар. Тогда превышение согласованности над несогласованностью есть*)

$$T = S - R = \sum_{i<j} \operatorname{sign}(\xi_j-\xi_i)\operatorname{sign}(\eta_j-\eta_i).$$

Значения T изменяются от $-n(n-1)/2$ до $n(n-1)/2$. Например, $\max T = n(n-1)/2$ достигается при идеальном согласии порядка $\xi_1,\ldots,\xi_n$ и $\eta_1,\ldots,\eta_n$. Для измерения степени согласия Кендэл предложил коэффициент

$$\tau = \frac{T}{\max T} = \frac{2T}{n(n-1)} = \frac{2(S-R)}{n(n-1)} = 1 - \frac{4}{n(n-1)}\,R, \qquad (23)$$

так как $S+R = n(n-1)/2$. Заметим, что величина R — это *количество инверсий* (см. пример 2 гл. 7), образованных величинами η_i, расположенными в порядке возрастания соответствующих ξ_i. Таким образом, коэффициент τ (линейно связанный с R) можно считать *мерой неупорядоченности* второй последовательности относительно первой.

Ввиду формулы (23) и асимптотической нормальности статистики R при справедливости гипотезы H_0 имеет место сходимость

$$\tau/\sqrt{\mathbf{D}\tau} \xrightarrow{d} Z \sim \mathcal{N}(0,1) \quad \text{при } n \to \infty,$$

где $\mathbf{D}\tau = 2(2n+5)/[9n(n-1)]$.

Обсудим *связь между коэффициентами* τ и ρ_S. Очевидно, статистика T представляется также в ранговой форме:

$$T = \sum_{i<j} \operatorname{sign}(R_j-R_i)\operatorname{sign}(S_j-S_i) = \sum_{i<j} \operatorname{sign}(T_j-T_i). \qquad (24)$$

*) Предполагается, что среди ξ_i и среди η_i нет совпадений.

Аналогично, $R = \sum_{i<j} I_{\{T_i > T_j\}}$. С учетом соотношения (23) получаем, что

$$\tau = 1 - \frac{4}{n^2 - n} \sum_{i<j} I_{\{T_i > T_j\}}.$$

Согласно задаче 5 для коэффициента ρ_S верна похожая формула:

$$\rho_S = 1 - \frac{12}{n^3 - n} \sum_{i<j} (j - i)\, I_{\{T_i > T_j\}}, \tag{25}$$

показывающая, что в случае ρ_S инверсиям придаются дополнительные веса $(j - i)$. Из-за этого возникает предположение, что ρ_S сильнее реагирует на несогласие ранжировок, чем τ. Однако М. Кендэл и А. Стьюарт в [35, с. 683] отмечают, что величины ρ_S и τ при справедливости гипотезы H_0 *сильно коррелированы*: коэффициент корреляции между ними равен $2\,(n+1)\big/\sqrt{2n(2n+5)}$. Он убывает от 1 при $n = 2$ до 0,98 при $n = 5$ и далее возрастает до 1 при $n \to \infty$.

Замечание 10. ***Обобщенный коэффициент корреляции*** [36]. Для удобства реализации на компьютере системы алгоритмов корреляционного анализа полезно вывести *обобщенную формулу* для вычисления разных парных корреляционных характеристик (таких, как τ, ρ_S и $\widehat{\rho}$, где

$$\widehat{\rho} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \Bigg/ \left[\sum_{i=1}^{n} (X_i - \overline{X})^2 \cdot \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \right]^{1/2}$$

обозначает *обычный коэффициент корреляции* между выборками $\boldsymbol{X} = (X_1, \ldots, X_n)$ и $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$). С этой целью определим некоторое правило, в соответствии с которым каждой паре (X_i, X_j) компонент вектора $\boldsymbol{X}$ приписывается число («метка») $c_{ij} = c_{ij}(\boldsymbol{X})$, причем это правило будет обладать свойством *отрицательной симметричности*: $c_{ij} = -c_{ji}$, $c_{ii} = 0$. Тогда *обобщенный коэффициент корреляции* между $\boldsymbol{X}$ и $\boldsymbol{Y}$ определяется формулой

$$\widehat{r} = \sum_{i<j} c_{ij}(\boldsymbol{X})\, c_{ij}(\boldsymbol{Y}) \Bigg/ \left[\sum_{i<j} c_{ij}^2(\boldsymbol{X}) \cdot \sum_{i<j} c_{ij}^2(\boldsymbol{Y}) \right]^{1/2}.$$

Убедимся, что коэффициенты $\widehat{\rho}$, ρ_S и τ могут быть получены как **частные случаи** обобщенного коэффициента $\widehat{r}$ при соответствующем выборе правила приписывания числовых «меток» c_{ij}.

1) Установим для любых $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$ справедливость тождества

$$\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n} \sum_{i<j} (X_j - X_i)(Y_j - Y_i).$$

При $X_i = Y_i$ оно следует из *теоремы о межточечных расстояниях*, доказанной при решении задачи 5 гл. 16 ($m_i = 1$, $m = n$). В общем случае оно выводится из указанной теоремы с помощью представления $AB = [(A + B)/2]^2 - [(A - B)/2]^2$ (см. задачу 6).

Ввиду установленного тождества при выборе в качестве «меток» $c_{ij}(\boldsymbol{X}) = X_j - X_i$ коэффициент $\widehat{r}$ преобразуется в $\widehat{\rho}$.

2) Положим $c_{ij}(\boldsymbol{X}) = R_j - R_i$, где R_i — ранг X_i в выборке $\boldsymbol{X}$. С учетом предыдущих рассуждений и определения коэффициента Спирмена (20) видим, что $\widehat{r}$ в этом случае совпадает с ρ_S.

3) Пусть $c_{ij}(\boldsymbol{X}) = \operatorname{sign}(X_j - X_i) = \operatorname{sign}(R_j - R_i)$. Тогда делимым в формуле для $\widehat{r}$ служит определенная выше статистика T, а делитель равен $n(n-1)/2$. Принимая во внимание формулу (23), заключаем, что обобщенный коэффициент $\widehat{r}$ превращается в τ.

§ 8. МНОЖЕСТВЕННАЯ И ЧАСТНАЯ КОРРЕЛЯЦИИ

Может представлять интерес задача измерения статистической связи сразу между $k \geqslant 3$ выборками. С этой целью Кендэлом был предложен ранговый *коэффициент конкордации (согласованности)*

$$W = \frac{12}{k^2(n^3 - n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} R_{ij} - \frac{k(n+1)}{2} \right)^2,$$

где R_{ij} — ранг (от 1 до n) i-го элемента в j-й выборке (столбце).

Укажем **некоторые свойства** коэффициента W (см. [36, гл. 6]).

1) $0 \leqslant W \leqslant 1$, причем $W = 1$ тогда и только тогда, когда все k ранжировок совпадают. То, что W не принимает отрицательных значений, объясняется тем обстоятельством, что в отличие от случая парных связей для $k \geqslant 3$ выборок противоположность согласованности утрачивается: упорядочения могут полностью совпадать, но не могут полностью не совпадать.

2) Обозначим через $\overline{\rho}_S$ *среднее арифметическое коэффициентов Спирмена по всем $k(k-1)/2$ парам выборок.* Тогда

$$W = [(k-1)\overline{\rho}_S + 1]/k.$$

Таким образом, W и $\overline{\rho}_S$ линейно связаны. В частности, при $k = 2$ имеем $W = (\rho_S + 1)/2$, т. е. коэффициент конкордации W линейно зависит от коэффициента Спирмена ρ_S.

3) Сравнение с критерием Фридмана из § 2 гл. 17 (с точностью до замены обозначений $k \rightleftarrows n$) показывает, что при больших n статистика $k(n-1)W$ распределена приближенно по закону хи-квадрат с $(n-1)$ степенями свободы.

Вопрос 3.
Будет ли значение $W = 0{,}09$ значимо велико на уровне 5% при $k = 20$ и $n = 15$?
(Воспользуйтесь табл. Т3 критических значений χ^2-распределения.)

Перейдем теперь к обсуждению понятия *частной или «очищенной» корреляции.* Начнем с примера из [2, с. 64].

«Даже если удалось установить тесную зависимость между двумя исследуемыми величинами, отсюда еще непосредственно не следует их *причинная взаимообусловленность.* Например, при анализе большого числа наблюдений, относящихся к отливке труб на сталелитейных заводах, была установлена положительная корреляционная связь между временем плавки и процентом забракованных труб [3]. Дать какое-либо причинное истолкование этой стохастической связи было невозможно, а поэтому рекомендации ограничить продолжительность плавки для снижения процента забракованных труб выглядели малосостоятельными. Действительно, спустя несколько лет обнаружили, что большая продолжительность плавки всегда была связана с использованием сырья специального состава. Этот вид сырья приводил одновременно к длительному времени плавки и большому проценту брака, хотя оба этих фактора взаимно независимы.

Таким образом, высокий коэффициент корреляции между продолжительностью плавки и процентом забракованных труб полностью обусловливался влиянием третьего, не учтенного при исследовании фактора — характеристики качества сырья. Если же этот фактор был бы с самого начала учтен, то никакой значимой корреляционной связи между временем плавки и процентом забракованных труб мы бы не обнаружили. За счет подобных эффектов (одновременного влияния неучтенных факторов на исследуемые переменные) может искажаться и смысл истинной связи между переменными, т. е., например, подсчеты приводят к положительному значению парного коэффициента корреляции, в то время как истинная связь между ними имеет отрицательный смысл. Такую корреляцию между двумя переменными часто называют «ложной». Более детально подобные ситуации — обнаружение и исключение «общих причинных факторов», расчет «очищенных», или *частных*, коэффициентов корреляции и т. п. — исследуют методами многомерного корреляционного анализа.»

Определение. Частным коэффициентом корреляции между случайными величинами X и Y при исключении влияния случайной величины Z называется

$$\rho(X,Y|Z) \equiv \rho_{XY|Z} = \frac{\rho(X,Y) - \rho(X,Z)\rho(Y,Z)}{\sqrt{(1-\rho^2(X,Z))(1-\rho^2(Y,Z))}}.$$

К этой формуле приводит попытка исключить зависимость от Z, заменив X и Y такими случайными величинами

$$X' = X - aZ, \quad Y' = Y - bZ,$$

которые некоррелированы с Z: $\rho(X',Z) = 0$ и $\rho(Y',Z) = 0$. Тогда «оставшаяся» корреляция представляет собой обычную корреляцию между X' и Y'.

Доказательство. Допустим для простоты, что $\mathsf{M}X = \mathsf{M}Y = \mathsf{M}Z = 0$. Для краткости введем обозначения $\sigma_\xi = \sqrt{\mathsf{D}\xi}$ и $\rho_{\xi\eta} = \rho(\xi,\eta)$. Константы a и b нужно выбрать так, чтобы имели место равенства

$$\mathsf{M}X'Z = \mathsf{M}XZ - a\mathsf{M}Z^2 = 0, \quad \mathsf{M}Y'Z = \mathsf{M}YZ - b\mathsf{M}Z^2 = 0.$$

Отсюда находим

$$a = \frac{\rho_{XZ}\sigma_X\sigma_Z}{\sigma_Z^2} = \rho_{XZ}\frac{\sigma_X}{\sigma_Z}, \quad b = \frac{\rho_{YZ}\sigma_Y\sigma_Z}{\sigma_Z^2} = \rho_{YZ}\frac{\sigma_Y}{\sigma_Z}. \tag{26}$$

Запишем обычный коэффициент корреляции между X' и Y':

$$\rho_{X'Y'} = \frac{\mathbf{M}(X-aZ)(Y-bZ)}{\sigma_{X-aZ}\,\sigma_{Y-bZ}}. \tag{27}$$

Числитель в формуле (27) можно представить в следующем виде:

$$\mathbf{M}XY - a\,\mathbf{M}YZ - b\,\mathbf{M}XZ + ab\,\mathbf{M}Z^2 =$$
$$= \rho_{XY}\,\sigma_X\,\sigma_Y - a\,\rho_{YZ}\,\sigma_Y\,\sigma_Z - b\,\rho_{XZ}\,\sigma_X\,\sigma_Z + ab\,\sigma_Z^2.$$

Заменив a и b в этом равенстве их значениями из соотношений (26), получим

$$\mathbf{M}(X-aZ)(Y-bZ) = (\rho_{XY} - \rho_{XZ}\,\rho_{YZ})\,\sigma_X\sigma_Y.$$

Точно также вычисляются дисперсии

$$\sigma^2_{X-aZ} = \mathbf{M}(X-aZ)^2 = (1-\rho^2_{XZ})\,\sigma^2_X,$$
$$\sigma^2_{Y-bZ} = \mathbf{M}(Y-bZ)^2 = (1-\rho^2_{YZ})\,\sigma^2_Y.$$

Подстановка всех этих выражений в формулу (27) приводит к равенству

$$\rho_{X'Y'} = \frac{\rho_{XY} - \rho_{XZ}\,\rho_{YZ}}{\sqrt{(1-\rho^2_{XZ})(1-\rho^2_{YZ})}} = \rho_{XY|Z}, \tag{28}$$

которое и требовалось установить. ■

Для получения оценки $\widehat{\rho}_{\boldsymbol{xy}|\boldsymbol{z}}$*) для коэффициента $\rho_{XY|Z}$ надо заменить в соотношении (28) теоретические коэффициенты корреляции выборочными (см. определение (20)):

$$\widehat{\rho}_{\boldsymbol{xy}|\boldsymbol{z}} = \frac{\widehat{\rho}_{\boldsymbol{xy}} - \widehat{\rho}_{\boldsymbol{xz}}\widehat{\rho}_{\boldsymbol{yz}}}{\sqrt{(1-\widehat{\rho}^2_{\boldsymbol{xz}})(1-\widehat{\rho}^2_{\boldsymbol{yz}})}}. \tag{29}$$

К этой формуле можно прийти точно так же, как и выше, отталкиваясь от условия ортогональности реализации выборки $\boldsymbol{z}$ и линейных комбинаций $\boldsymbol{x}' = \boldsymbol{x} - a\boldsymbol{z}$ и $\boldsymbol{y}' = \boldsymbol{y} - b\boldsymbol{z}$. В этой интерпретации $\widehat{\rho}_{\boldsymbol{xy}|\boldsymbol{z}}$ представляет собой косинус угла между проекциями векторов $\boldsymbol{x}$ и $\boldsymbol{y}$ на подпространство в $\mathbb{R}^n$, ортогональное вектору $\boldsymbol{z}$ (рис. 26).

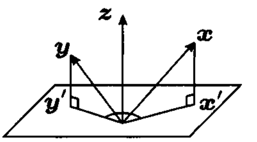


Рис. 26

Если дополнительно предположить, что $\boldsymbol{X}$, $\boldsymbol{Y}$ и $\boldsymbol{Z}$ — выборки из *независимых нормальных законов*, то (как доказано в [13, с. 370]) выборочный частный коэффициент корреляции $\widehat{\rho}_{\boldsymbol{XY}|\boldsymbol{Z}}$ будет распределен точно также, как и обычный выборочный коэффициент $\widehat{\rho}_{\boldsymbol{XY}}$, но для выборок размера не n, а $n-1$. Отсюда и из задачи 6 гл. 11 вытекает сходимость распределения случайной величины $\sqrt{n}\ \operatorname{arcth}\widehat{\rho}_{\boldsymbol{XY}|\boldsymbol{Z}}$ к закону $\mathcal{N}(0,1)$ при $n \to \infty$.

Пример 10 ([2, с. 85]). По итогам года у 37 однородных предприятий легкой промышленности были зарегистрированы следующие (среднемесячные) показатели их работы: $\boldsymbol{x}$ — значения характеристики качества ткани (в баллах), $\boldsymbol{y}$ — количества профилактических наладок автоматической линии, $\boldsymbol{z}$ — случаи обрывов нити.

*) Здесь неслучайный вектор $\boldsymbol{x} = (x_1, \ldots, x_n)$ обозначает реализацию выборки $\boldsymbol{X} = (X_1, \ldots, X_n)$, где X_i независимы и распределены так же, как и случайная величина X.

На основе этих данных были подсчитаны парные коэффициенты корреляции: $\widehat{\rho}_{xy} = 0{,}105$, $\widehat{\rho}_{xz} = 0{,}024$, $\widehat{\rho}_{yz} = 0{,}996$. Проверка на статистическую значимость свидетельствует об отсутствии связи между качеством ткани, с одной стороны, и числом профилактических наладок и обрывов нити — с другой, что не согласуется с профессиональными представлениями технолога.

Однако расчет *частных* коэффициентов корреляции по формуле (29) дает значения $\widehat{\rho}_{xy|z} = 0{,}908$ и $\widehat{\rho}_{xz|y} = -0{,}907$, которые вполне соответствуют представлению о естественном характере связей между изучаемыми показателями.

В заключение отметим, что ранговый коэффициент Кендэла τ (в отличие от коэффициента Спирмена ρ_S) переносится на случай частной корреляции с помощью формулы, аналогичной формуле (29):

$$\tau_{xy|z} = \frac{\tau_{xy} - \tau_{xz}\tau_{yz}}{\sqrt{(1-\tau_{xz}^2)(1-\tau_{yz}^2)}}$$

(см. [36, гл. 8]). Критерии значимости для $\tau_{xy|z}$ можно отыскать в журнальных статьях, указанных на с. 216 книги [86].

§ 9. ТАБЛИЦЫ СОПРЯЖЕННОСТИ

Рассмотрим задачу выявления статистической связи для сгруппированных (разбитых на категории) данных. Если ранее обсуждались случаи *количественных* (§§ 1–6) и *порядковых (ранговых)* (§§ 7–8) переменных, то теперь переменные — качественные и описываются номером группы, а данные представлены в виде *таблицы сопряженности (признаков)* $\|\nu_{ij}\|_{n\times m}$, в которой ν_{ij} — числа объектов с признаками i, j.*)

Опишем **три выборочные схемы**, приводящие к таблицам сопряженности ([2, с. 125]).

Схема I возникает в случае, когда строки $(\nu_{i1},\ldots,\nu_{im})$ таблицы данных $(i = 1,\ldots,n)$ можно рассматривать как независимые выборки из полиномиальных распределений (см. § 5 гл. 10 и доказательство теоремы 1 гл. 18) с вероятностями q_{ij} $(\sum_{j=1}^{m} q_{ij} = 1)$ и заданными числами наблюдений $n_i = \sum_{j=1}^{m} \nu_{ij}$. Такая организация данных обычно возникает, когда хотят сравнить между собой несколько одномерных распределений, представленных выборками заранее заданного размера. Наиболее важной для схемы I является *гипотеза однородности*

$$H_I\colon q_{ij} = q_{\cdot j}, \quad \text{где } q_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} q_{ij},$$

*) Таблицы с тремя и более входами анализируются в книгах [7], [47].

которая подробно обсуждалась ранее в § 3 гл. 18.

Схема II. Предполагается, что $(\nu_{11}, \dots, \nu_{nm})$ имеют полиномиальное распределение с вероятностями $(p_{11}, \dots, p_{nm})$ и фиксированным общим числом наблюдений $N = \sum\limits_{i,j} \nu_{ij}$. Таблица сопряженности в этом случае является обычной двумерной гистограммой для N наблюдений. Гипотезе H_I из схемы I в схеме II соответствует *гипотеза независимости*, состоящая в том, что совместное распределение есть произведение маргинальных (частных) распределений:

$$H_{II}\colon p_{ij} = r_i s_j, \quad \text{где } r_i = \sum_{j=1}^{m} p_{ij}, \quad s_j = \sum_{i=1}^{n} p_{ij}.$$

Схема III возникает, когда в схеме II общее число наблюдений рассматривается как случайная величина. Ее важным частным случаем является случай, когда все ν_{ij} независимы и распределены по закону Пуассона с параметрами λ_{ij}. Тогда их сумма N также имеет распределение Пуассона с параметром $c = \sum\limits_{i,j} \lambda_{ij}$ (см. задачу 3 гл. 10). Гипотезам H_I и H_{II} соответствует *гипотеза мультипликативности*

$$H_{III}\colon \lambda_{ij} = a_i b_j / c, \quad \text{где } a_i = \sum_{j=1}^{m} \lambda_{ij}, \quad b_j = \sum_{i=1}^{n} \lambda_{ij}.$$

В качестве примера применения схемы III может быть рассмотрена задача, в которой ν_{ij} — число отказов (аварий) i-го вида на установках j-го типа в течение заданного времени наблюдения. Параметры λ_{ij} отражают ожидаемые количества отказов.

Можно доказать, что если в схеме III зафиксировать N, то она переходит в схему II с $p_{ij} = \lambda_{ij}/c$. При этом гипотеза H_{III} преобразуется в гипотезу H_{II}. Аналогично, если зафиксировать в схеме II суммы по строкам $n_i = \nu_{i1} + \ldots + \nu_{im}$, то схема II переходит в схему I с $q_{ij} = p_{ij}/r_i$, а гипотеза H_{II} — в гипотезу H_I.

Для проверки гипотез H_I—H_{III} применяется вариант **критерия хи-квадрат**, статистика которого имеет вид

$$X^2 = N \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(\nu_{ij} - n_i m_j / N)^2}{n_i m_j}, \text{ где } n_i = \sum_{j=1}^{m} \nu_{ij}, \; m_j = \sum_{i=1}^{n} \nu_{ij}.$$

При справедливости проверяемой гипотезы при достаточно больших N статистика X^2 приближенно распределена по закону хи-квадрат с $(n-1)(m-1)$ степенями свободы.

Для схемы II это утверждение вытекает из теоремы Фишера (см. формулу (6) гл. 18). В этом случае имеется $(m + n - 2)$ неизвестных параметров $r_1, \dots, r_{n-1}, s_1, \dots, s_{m-1}$. Методом Лагранжа находим, что оценками максимального правдоподобия для них будут величины $\widehat{r}_i = n_i/N$ и $\widehat{s}_j = m_j/N$ (задача 7). При этом число степеней свободы предельного закона хи-квадрат равно $nm - 1 - (m + n - 2) = (n-1)(m-1)$.

Пример 11 ([44, с. 481]). В приведенной ниже таблице представлены результаты социологического обследования о связи между доходом семей и количеством детей в них. Признак A означает количество детей и принимает значения 0, 1, 2, 3, $\geqslant 4$. Признак B указывает, какому из диапазонов (0–1), (1–2), (2–3), ($\geqslant 3$) (за единицу принято 1000 шведских крон) принадлежит доход семьи.

$A \backslash B$	0–1	1–2	2–3	$\geqslant 3$	n_i
0	2161	3577	2184	1636	**9558**
1	2755	5081	2222	1052	**11110**
2	936	1753	640	306	**3635**
3	225	419	96	38	**778**
$\geqslant 4$	39	98	31	14	**182**
m_j	**6116**	**10928**	**5173**	**3046**	**25263**

Значение статистики X^2 равно 568,6, что значимо больше критической границы 32,9 уровня 0,001 закона хи-квадрат с $(5-1)(4-1) = 12$ степенями свободы (см. табл. Т3). Поэтому гипотеза независимости признаков A и B отвергается.*)

ЗАДАЧИ

Помучишься — так научишься.

1. Пусть $\boldsymbol{A}$, $\boldsymbol{B}$ и $\boldsymbol{A}-\boldsymbol{B}$ — неотрицательно определенные матрицы. Докажите, что

   а) $\operatorname{tr}\boldsymbol{A} \geqslant \operatorname{tr}\boldsymbol{B}$; $\operatorname{tr}\boldsymbol{A} = \operatorname{tr}\boldsymbol{B} \iff \boldsymbol{A} = \boldsymbol{B}$,

   Указание. Используйте приведение $\boldsymbol{A}-\boldsymbol{B}$ к главным осям.

   б) $\det\boldsymbol{A} \geqslant \det\boldsymbol{B}$; если $\det\boldsymbol{B} > 0$, то $\det\boldsymbol{A} = \det\boldsymbol{B} \iff \boldsymbol{A} = \boldsymbol{B}$.

   Указание. Рассмотрите сначала случай $\boldsymbol{B} = \boldsymbol{E}$ и докажите, что тогда все собственные значения матрицы $\boldsymbol{A}-\boldsymbol{E}$ не меньше 1.

2.* Выведите соотношения (19).

3. Получите представление (22) рангового коэффициента Спирмена ρ_S.

   Указание. Разложите $\sum\left[\left(i-\frac{n+1}{2}\right)-\left(T_i-\frac{n+1}{2}\right)\right]^2$.

4.* Докажите, что $\mathbf{D}\rho_S = 1/(n-1)$ при справедливости гипотезы H_0.

   Указание. Положим $\xi_i = R_i - \dfrac{n+1}{2}$. В силу равенства $\sum\xi_i = 0$ имеем

   $$0 = \mathbf{M}\left(\xi_1\sum\xi_i\right) = \mathbf{M}\xi_1^2 + (n-1)\mathbf{M}\xi_1\xi_2.$$

5.* Проверьте справедливость для ρ_S представления (25).

6. Проверьте, что для любых $x_1,\ldots,x_n$ и $y_1,\ldots,y_n$ верно тождество

   $$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \frac{1}{n}\sum_{i<j}(x_j-x_i)(y_j-y_i).$$

---

*) Однако в [11, с. 412] отмечено, что более тонкий анализ этих данных указывает на очень слабую зависимость между A и B.

Задачи

Обозначим через R_i номер в порядке возрастания (*ранг*) элемента X_i из выборки $X_1, \ldots, X_n$, через S_i — ранг элемента Y_i из выборки $Y_1, \ldots, Y_n$. Переставив пары (R_i, S_i) в порядке возрастания первой компоненты, получим набор $(1, T_1), \ldots, (n, T_n)$.

1. Вычислите математическое ожидание и дисперсию случайной величины R_i.

2. Найдите $\mathsf{P}(R_i = k, R_j = l)$ при $i \neq j, k \neq l$ и ковариацию $\mathbf{cov}(R_i, R_j)$ при $i \neq j$.

3. Докажите, что коэффициент ранговой корреляции Спирмена совпадает с коэффициентом Пирсона, вычисленным на основе рангов R_i и S_i.

4. Докажите, что для коэффициента ранговой корреляции Спирмена $\hat{\rho}_S$ верна формула

$$\hat{\rho}_S = 1 - \frac{12}{n^3 - n} \sum_{i<j} (j - i) I_{\{T_i > T_j\}}.$$

5.* Докажите, что при выполнении гипотезы независимости признаков дисперсия $\mathsf{D}\hat{\rho}_S = 1/(n-1)$.