МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

РЯЗАНСКИЙ ГОСУДАРСТВЕННЫЙ РАДИОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Е.П. ЧУРАКОВ

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Рязань 2013

Министерство образования и науки Российской Федерации

Рязанский государственный радиотехнический университет

Е.П. ЧУРАКОВ

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Часть 1

Учебное пособие

Рязань 2013

УДК 519.237(075.8)
ББК 22.172я73

Многомерные статистические методы часть 1: учеб. пособие / Е.П. Чураков: Рязан. гос. радиотехн. ун-т. Рязань, 2013. 96 с.

Систематизированы основные сведения по многомерным статистическим методам (многомерному статистическому анализу). Кратко изложены основы математико-вероятностного аппарата, применяемого при многомерном анализе. Значительно большее внимание уделено корреляционному анализу многомерных случайных величин, включая вопросы ранговой корреляции и дисперсионного анализа. Изложены основные подходы к преодолению проблемы избыточной размерности многомерных задач, использующие концепции компонентного и факторного анализа.

Предназначено для студентов специальности 080116 “Математические методы в экономике” и направлений 010400 “Прикладная математика и информатика” с профилем подготовки “Математическое и информационное обеспечение экономической деятельности” и 080500 “Бизнес-информатика”.

Библиогр.: 21 назв.

Многомерный анализ, векторные случайные величины, ранговая корреляция, компонентный и факторный анализ

Печатается по решению редакционно-издательского совета Рязанского государственного радиотехнического университета.

Рецензент: кафедра ЭиММ Рязанского государственного радиотехнического университета (зав. кафедрой В.К. Клочко)

© Рязанский государственный
радиотехнический университет, 2013

Введение

Многомерные статистические методы (многомерный статистический анализ) представляют собой углубленный (или, как сейчас говорят, продвинутый) курс математической статистики, направленный на изучение свойств некоторой многомерной случайной величины

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_k \end{bmatrix} \in R^k ,$$

состоящей из k упорядоченных в форме k-мерного вектора-столбца случайных скалярных компонентов $X_1, X_2, \dots, X_k$ и принадлежащей k-мерному координатному пространству $\boldsymbol{R}^k$. Целью многомерного анализа являются сбор статистических (экспериментальных) данных об этом векторе, предварительный анализ этих данных и их аналитическая обработка с целью установления вероятностных свойств как отдельных компонентов вектора $\boldsymbol{X}$, так и вектора $\boldsymbol{X}$ в целом как единого математического объекта, объединяющего k случайных величин. В результате обработки экспериментальных данных (наблюдений) должны появиться отсутствующие сведения о вероятностных свойствах отдельных компонентов вектора $\boldsymbol{X}$, характере взаимосвязей их друг с другом, о выявлении скрытых (латентных) факторов, влияющих на вектор $\boldsymbol{X}$, о возможности уменьшения размерности k вектора $\boldsymbol{X}$ без ущерба для вероятностного восприятия этого вектора.

Возможны ситуации, в которых изучаемое явление характеризуется многими аналогичными $\boldsymbol{X}$ векторами: $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \dots, \boldsymbol{X}^{(m)}$. Тогда, помимо перечисленных, может возникнуть дополнительная проблема разбиения всей совокупности объектов (векторов) на некоторое известное или неизвестное число групп (классов), состоящих только из однородных по некоторому признаку объектов. Интересной может оказаться задача отнесения какого-либо вектора к одному из заданных семейств векторов.

Перечисленное незначительное число задач многомерного анализа не должно создать иллюзию завершенности этого перечня. Трудно (а быть может, и невозможно) перечислить все проблемы, попадающие в прокрустово ложе многомерного анализа. Реальная действительность порождает все новые и новые проблемы, для преодоления которых

привлекаются как уже известные методы многомерного анализа, так и вновь создаваемые под эти проблемы. Многомерные методы – это своеобразный динамично развивающийся научный организм, и настоящее пособие представляет собой только визитную карточку этого организма, способствующую, по мнению автора, последующей плодотворной дружбе читателя и многомерных методов.

Список литературных источников по многомерным методам и частным вопросам многомерного анализа достаточно внушителен. Сошлемся только на работы относительно недавнего времени [1-5], в которых приводятся ссылки и на более ранние издания. Помимо этого, известны многочисленные работы "корпоративного" – внутривузовского характера, например [6-10], "добраться" до которых можно с помощью Internet. Важно заметить, что задачи многомерного анализа успешно решаются только с применением вычислительных средств. Поэтому большинство современных пакетов прикладных программ предусматривает в своем составе многочисленные встроенные функции для проведения многомерного анализа, культура плодотворного общения с которыми требует соответствующей культуры в познании математического аппарата решения соответствующих задач.

Трудности с приобретением литературы по многомерным методам, обусловленные, прежде всего, ограниченными в последнее время тиражами научных изданий, заставили автора взяться за подготовку данного учебного пособия. Оно ориентировано на студентов специальности "Математические методы в экономике" и направления "Прикладная математика и информатика" с профилем подготовки "Математическое и информационное обеспечение экономической деятельности", но, по мнению автора, может быть полезно и для иных направлений подготовки, учебными планами которых предусмотрена подобная дисциплина. Применяемый в пособии математико-вероятностный аппарат излагается "инженерным" языком в стиле замечательной книги Е. С. Венцель [11], лекции которой, прослушанные автором данного пособия в начале 60-х годов прошлого века на факультете переподготовки дипломированных инженеров в МГУ, оставили яркий след и породили уважение и пристрастие к вероятностным задачам. Существует надежда, что и читателям пособия, склонным к использованию аксиоматико-вероятностного стиля изложения [12], этот вариант не покажется архаичным, и в любом случае все их замечания будут восприняты с благодарностью.

1. ХАРАКТЕРИСТИКИ И КОРРЕЛЯЦИОННЫЙ АНАЛИЗ МНОГОМЕРНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

1.1. Скалярные случайные величины

Основным объектом исследования в многомерном анализе, что-уже отмечалось во введении, является векторная *k*-мерная случайная величина $\boldsymbol{X}$ с компонентами X_i, $i = \overline{1,k}$. Прежде чем приступить к описанию ее свойств, целесообразно напомнить основные вероятностные характеристики ее скалярных компонентов. Хотя эта процедура является прерогативой традиционного курса теории вероятностей и ее содержание, по идее, читателю известно, кратко напомним существо соответствующих построений, помня о роли повторения в процессе обучения.

Определение 1.1. Пусть проводится некоторый эксперимент (наблюдение), предполагаемый исход которого характеризуется непрерывной величиной *X*. Эту величину принято называть *случайной,* если до проведения эксперимента ее точное значение указать нельзя, но можно указать вероятность выполнения любого события вида $X < x$, где *x* – произвольное число из диапазона $(-\infty, \infty)$, т.е. вероятность того, что величина *X* окажется меньше любого числа *x*. (Заметим, что в ряде работ это неравенство задается как нестрогое $X \le x$.

Определение 1.2. Функцию $F_X(x) = P(X < x)$, где $P(X < x)$ - вероятность выполнения события $X < x$, называют *одномерным законом распределения вероятностей* случайной величины *X* (функцией распределения вероятностей).

Напомним некоторые характерные свойства этой функции:

1) $F_X(-\infty) = 0$;

2) $F_X(\infty) = 1$;

3) $F_X(b) - F_X(a) = P(a \le X < b)$, где $b>a$ и $a,b \in R$;

4) $F_X(b) \ge F_X(a)$ при $b > a$;

5) функция $F_X(x)$ непрерывна слева, т.е.

$$F_X(x_0) = \lim_{x \to x_0 - 0} F_X(x) \text{ и } P(X < x_0) = F(x_0);$$

6) $P(X = a) = 0$, если функция $F_X(x)$ в точке *a* непрерывна, и

$P(X = a) = \lim\limits_{b \to a+0} F_x(b) - F_x(a)$,

если в этой точке имеется разрыв первого рода.

Функция распределения вероятностей, таким образом, является неубывающей непрерывной слева с множеством значений на отрезке [0,1]; она в общем случае имеет участки монотонного возрастания, участки постоянства и в некоторых точках может иметь разрывы первого рода. Но в большинстве реальных задач эта функция оказывается непрерывной.

С функцией $F_X(x)$ связывают два важных понятия, широко используемых при рассмотрении последующего материала.

Определение 1.3. Пусть $q \in [0,1]$. Тогда корень уравнения $F_X(x) = q$ называется *квантилем* (иногда *квантилью*) порядка q распределения $F_X(x)$ и часто обозначается символом u_q. Иными словами, квантиль удовлетворяет условию $P(X < u_q) = q$.

Определение 1.4. Пусть $Q \in [0,100]$. Тогда корень уравнения $1 - F_X(x) = 10^{-2} Q$ называется *Q-процентной точкой* функции $F_X(x)$ и часто обозначается символом w_Q. Иными словами, *Q*-процентная точка удовлетворяет условию $1 - P(X < w_Q) = 10^{-2} Q$.

Квантили и процентные точки тесно связаны друг с другом. Действительно, если u_q – квантиль *q*-го порядка, то эта же величина будет 10^2 (1-*q*)-процентной точкой, т.е. очевидна связь:

$$u_q = w_{10^2(1-q)},\ w_Q = u_{1-10^{-2}Q}.$$

Квантилю, например, порядка 0,1 будет соответствовать 90-процентная точка, а 40-процентная точка будет совпадать с квантилем порядка 0,6. Поэтому практически приходится вычислять только один вид этих точек, а именно – квантили. Они содержатся в обширных таблицах, приводимых в многочисленных литературных источниках. Однако табличные источники информации – это дань прошлому. Большинство современных пакетов прикладных программ предусматривают специальные встроенные функции, которые возвращают квантили произвольных порядков для широкого спектра функций $F_X(x)$.

Определение 1.5. Пусть существует функция $f_X(x)$ такая, что можно представить

$$\int_{-\infty}^{x} f_X(x)dx = F_X(x) \qquad (1.1)$$

или при дифференцируемой функции $F_X(x)$

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Тогда функция $f_X(x)$ называется *одномерной плотностью вероятностей* случайной величины *X*.

Смысл плотности вероятностей заключается в том, что с точностью до бесконечно малых высших порядков выполняется равенство

$$f_X(x)dx = P\left(x \le X < x + dx\right)$$

и функция $f_X(x)$ показывает, как эта вероятность распределяется вдоль оси x. Будучи производной от неубывающей функции $F_X(x)$, плотность вероятностей является неотрицательной функцией, а вследствие (1.1) справедливы равенства

$$\int_{-\infty}^{\infty} f_X(x)dx = 1,$$

$$\int_{a}^{b} f_X(x)dx = F_X(b) - F_X(a) = P\left(a \le X < b\right),$$

из которых первое известно как условие *нормировки.*

Формально любая нормированная неотрицательная функция может претендовать на то, чтобы выступить в качестве плотности вероятностей какой-либо случайной величины. Однако реально используется ограниченный перечень таких функций. Укажем те из них, с которыми придется встречаться в последующем.

Гауссова или *нормально распределенная* случайная величина *X* имеет плотность вероятностей вида

$$f_X(x) = \frac{1}{\sqrt{2\pi D}} exp\left(-\frac{1}{2D}\left(x-m\right)^2\right), \qquad (1.2)$$

где D, m - параметры распределения, смысл которых обсудим позже. Так как эта плотность определяется двумя параметрами, то лаконично гауссову величину *X* принято отображать записью

$X \sim N(m, D)$. Если $m = 0, D = 1,$ то используют обозначение $X \sim N(0,1)$ и величину X в этом случае называют *стандартной* гауссовой величиной. Такие величины играют важную роль в статистическом анализе, так как с их помощью формируются другие случайные величины.

Пусть $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ - последовательность из n независимых (строгий смысл этого понятия уточним позже) стандартных гауссовых величин. Образуем из них новую случайную величину X по правилу

$$X = \sum_{i=1}^{n} \varepsilon_i^2 .$$

Тогда говорят, что величина X *распределена по закону* χ^2 (читается "хи квадрат") *с n степенями свободы,* и пишут $X \sim \chi^2(n)$.

Пусть $\varepsilon_0, \varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ - последовательность из $n+1$ независимых стандартных гауссовых величин. Образуем из них новую случайную величину Y по правилу

$$Y = \frac{\varepsilon_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2}}.$$

Случайная величина Y в этом случае называется *распределенной по закону Стьюдента с n степенями свободы.* Сокращенно пишут $Y \sim t(n)$.

Пусть имеются две последовательности независимых стандартных гауссовых величин: $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ и $\eta_1, \eta_2, ..., \eta_m$. Из них сформируем новую случайную величину

$$Z = \frac{\frac{1}{m}\sum_{i=1}^{m} \eta_i^2}{\frac{1}{n}\sum_{j=1}^{n} \varepsilon_j^2} .$$

В таком случае говорят, что случайная величина Z *распределена по закону Фишера с m степенями свободы числителя и n степенями свободы знаменателя*, и пишут $Z \sim F(m,n)$.

В книгах по теории вероятностей содержатся исчерпывающие характеристики этих случайных величин, но мы ограничимся только их определениями в терминах выражения через независимые гауссовы стандартные величины.

При решении многочисленных прикладных задач, требующих привлечения аппарата случайных величин, часто удается ограничиться достаточно скромной информацией о них, содержащейся в так называемых *числовых характеристиках случайных величин*. Напомним соответствующие понятия в объеме, необходимом для последующего изложения материала.

Определение 1.6. Пусть $\varphi(X)$ - некоторая неслучайная функция от случайной величины *X*. Тогда интегральную операцию вида $\int_{-\infty}^{\infty} \varphi(x) f_X(x)dx$, где $f_X(x)$ – плотность вероятностей *X*, называют *оператором усреднения или взятия математического ожидания,* а результат этой операции называют *средним значением* функции $\varphi\left(X\right)$.

Для лаконичного обозначения этой операции в последующем используем *символ усреднения M(.)* (оператор усреднения), т.е. обозначение

$$M\left(\varphi\left(X\right)\right) = \int_{-\infty}^{\infty} \varphi\left(x\right) f_X\left(x\right) dx .$$

Заметим, что оператор усреднения является *линейным* в смысле выполнения условия

$$M\left(\sum_{j=1}^{s} a_j \varphi_j\left(X\right)\right) = \sum_{j=1}^{s} a_j M\left(\varphi_j\left(X\right)\right),\ a_j = const,\ j = \overline{1,s},$$

т.е. математическое ожидание от линейной комбинации функций случайного аргумента равняется той же линейной комбинации математических ожиданий от функций. Это свойство диктуется интегральной природой оператора.

Определение 1.7. *Средним значением* или *математическим ожиданием* случайной величины *X* называют число

$$m_X = M\left(X\right) = \int_{-\infty}^{\infty} x f_X(x)dx . \qquad (1.3)$$

Величину X-m_X называют *центрированной* случайной величиной. Ее среднее значение равно нулю.

Определение 1.8. Среднее значение квадрата центрированной случайной величины X-m_X называют *дисперсией* D_X случайной величины X, т.е.

$$D_X = M(\left(X - m_X\right)^2) = \int_{-\infty}^{\infty} \left(x - m_X\right)^2 f_X(x)dx, \qquad (1.4)$$

а число $\sigma_X = \sqrt{D_X}$ называют *среднеквадратичным отклонением* величины X.

Обратим внимание на то, что m_X, D_X, σ_X - постоянные параметры случайной величины X.

Можно показать, что параметры m, D, фигурирующие в определении 1.2 нормальной плотности, представляют собой *математическое ожидание* m_X и *дисперсию* D_X гауссовой случайной величины. При этом дисперсия D_X является своеобразной мерой разброса возможных значений случайной величины относительно ее среднего значения m_X. Например, для математического ожидания имеем

$$m_X = \int_{-\infty}^{\infty} xf\left(x\right)dx = \frac{1}{\sqrt{2\pi D}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2D}\left(x-m\right)^2\right)dx .$$

Введем новую переменную $\frac{x-m}{\sqrt{D}} = t,\ dx = \sqrt{D}dt$ и получим

$$m_X = \frac{1}{\sqrt{2\pi}}\left(m\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}t^2\right)dt + \sqrt{D}\int_{-\infty}^{\infty} t\exp\left(-\frac{1}{2}t^2\right)dt\right) = m,$$

так как первый интеграл в этом выражении (интеграл Пуассона), будучи табличным, равняется $\sqrt{2\pi}$ (например, [13]), а второй равен нулю как интеграл от нечетной функции в симметричных пределах. Аналогичным образом доказывается $D = D_X$.

Известно так называемое "правило 3σ", которое утверждает, что приближенно выполняется неравенство

$$P(m_X - 3\sigma_X < X < m_X + 3\sigma_X) > 0.95 . \qquad (1.5)$$

Необходимо иметь в виду:
если $c=const$, то $m_c = M(c)=c$, $M(cX)=cm_X$, $Dc=0$, $D_{cX}=c^2D_X$;
если $c_1, c_2, \ldots, c_k=const$ и $X_1, X_2, \ldots, X_k$ – случайные величины, то

$$M(\sum_{i=1}^{k} c_i X_i) = \sum_{i=1}^{k} c_i m_i, \quad m_i = M(X_i),$$

а если случайные величины независимые, то

$$M\left(\left(\sum_{i=1}^{k}\left(c_i X_i - c_i m_i\right)\right)^2\right) = \sum_{i=1}^{k} c_i^2 D_i, \quad D_i = M\left(\left(X_i - m_i\right)^2\right).$$

1.2. Выборочные оценки числовых характеристик скалярных случайных величин

При решении многих прикладных задач удается ограничиться “скромной” информацией о случайной величине в объеме ее числовых характеристик (1.3), (1.4). Поэтому нужно уметь их вычислять, для чего, прежде всего, необходимо знать плотность вероятностей $f_X(x)$. Однако в большинстве реальных задач именно эта плотность неизвестна. Тогда стремятся обойтись без функции $f_X(x)$, получая различными способами вместо истинных характеристик m_X, D_X их приближенные значения $\hat{m}_X$, $\hat{D}_X$, которые мы будем называть *оценками* числовых характеристик. В качестве исходного материала для получения оценок, независимо от способа их формирования, используют результаты конкретных наблюдений за случайной величиной X, проводимых при одних и тех же условиях на различных объектах. Напомним некоторые сопутствующие ситуации понятия.

Теоретическое (гипотетическое) множество, состоящее из N, где в общем случае не исключается $N = \infty$, независимых случайных величин, имеющих один и тот же закон распределения вероятностей $F_X(x)$, называют *генеральной совокупностью*. Подмножество генеральной совокупности, состоящее из $n<N$ ее элементов, называют *выборкой*. Этот же термин используют и при регистрации конкретных значений, которые выборка принимает при проведении n наблюдений за случайной величиной X, и соответствующие результаты обозначают как x_1, x_2, … , x_n или сокращенно x_i, $i = \overline{1,n}$. Именно наблюдения x_i, $i = \overline{1,n}$, используются при поиске оценок $\hat{m}_X$, $\hat{D}_X$, т.е. эти оценки являются функциями проведенных наблюдений. Так как до проведения наблюдений элементы выборки случайны, то и оценки $\hat{m}_X$, $\hat{D}_X$ - случайные величины. Несмотря на многочисленность способов, средствами которых эти оценки могут быть вычислены, существует ряд “пожеланий”, к достижению которых необходимо стремиться. На примере абстрактной оценки $\hat{a}$ некоторого неизвестного параметра a напомним их.

Величину $\eta_a = \hat{a} - a$ называют *ошибкой оценивания*. Оценку $\hat{a}$ называют *несмещенной*, если среднее значение ее ошибки равно нулю, или, что то же самое, среднее значение оценки равняется оцениваемому параметру: $M(\eta_a) = 0 \Leftrightarrow M(\hat{a}) = a$. Оценку $\hat{a}$ называют *состоятельной*, если она при неограниченном увеличении объема выборки *n* сходится по вероятности к истинному значению параметра *a*, т.е. если при $\forall \varepsilon > 0$ выполняется условие

$$\lim_{n\to\infty} P\left(\left|\hat{a} - a\right| < \varepsilon\right) = 1.$$

Наконец, оценку называют *эффективной*, если средний квадрат ее ошибки не больше, чем средний квадрат ошибки любой другой оценки. При поиске оценки $\hat{a}$ следует стремиться к тому, чтобы она оказалась несмещенной, состоятельной и эффективной. Но при этом полезно помнить, что при конечном *n* и любом способе оценивания невозможно получить оценку со сколь угодно малым средним квадратом ошибки. Существует так называемое *неравенство Рао - Крамера*, которое снизу ограничивает средний квадрат ошибки (например, [14]).

Возвратимся к поиску оценок $\hat{m}_X, \hat{D}_X$. Наиболее часто с этой целью используется *метод максимального правдоподобия*, при котором, не строго говоря, максимизируется вероятность наблюдать величины x_i, $i = \overline{1,n}$. Ограничив рассмотрение гауссовым случаем (1.2), можем записать совместную плотность вероятностей $\phi(m_X, D_X)$ независимых случайных величин $X_1, X_2, \ldots, X_n$, образующих выборку из генеральной гауссовой совокупности,

$$\phi(m_X, D_X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi D_X}} exp\left(\frac{1}{2D_X}\left(x_i - m_X\right)^2\right) \qquad (1.6)$$

или после логарифмирования

$$\ln \phi(m_X, D_X) = -\frac{n}{2}\ln\left(2\pi D_X\right) - \frac{1}{2D_X}\sum_{i=1}^{n}\left(x_i - m_X\right)^2.$$

Найдем величины m_X и D_X, максимизирующие эту функцию. Необходимые условия максимума имеют вид

$$\frac{\partial}{\partial m_X}\ln\phi\left(m_X, D_X\right)=\frac{1}{D_X}\sum_{i=1}^{n}\left(x_i-m_X\right)=0,$$

$$\frac{\partial}{\partial D_X}\ln\phi\left(m_X, D_X\right)=-\frac{n}{2}D_X^{-1}+\frac{1}{2}D_X^{-2}\sum_{i=1}^{n}\left(x_i-m_X\right)^2=0.$$

Функция $\phi(m_X, D_X)$ является выпуклой. Поэтому необходимые условия являются одновременно и достаточными. Решения этих уравнений и принимаются за искомые *максимально правдоподобные* оценки, т.е.

$$\hat{m}_X=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \hat{D}_X=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\hat{m}_X\right)^2. \tag{1.7}$$

Исследуем смещение оценок. Для первой из них имеем

$$M\left(\hat{m}_X\right)=\frac{1}{n}\sum_{i=1}^{n}M\left(X_i\right)=\frac{1}{n}\sum_{i=1}^{n}m_X=m_X,$$

т.е. оценка $\hat{m}_X$ является *несмещенной*. Вторую оценку, раскрыв квадрат, представим в виде

$$\hat{D}_X=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\hat{m}^2$$

и, следовательно,

$$M\left(\hat{D}_X\right)=\frac{1}{n}\sum_{i=1}^{n}M\left(X_i^2\right)-M\left(\hat{m}_X^2\right).$$

Для каждого из слагаемых имеем:

$$M\left(X_i^2\right)=M\left(\left(X_i-m_X+m_X\right)^2\right)=D_X+m_X^2,$$

где использовано следствие центрированности $M\left(X_i-m_X\right)=0$;

$$M\left(\hat{m}_X^2\right)=M\left(\frac{1}{n^2}\left(\sum_{i=1}^{n}X_i\right)^2\right)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}M\left(X_iX_j\right).$$

Так как случайные величины X_i и X_j независимые, то

$$M\left(X_i X_j\right)=\begin{cases} D_X+m_X^2, & i=j, \\ m_X^2, & i\neq j. \end{cases}$$

Двойная сумма в выражении для $M\left(\widehat{m}_X^2\right)$ содержит n^2 слагаемых, из них в n слагаемых $i=j$ и в оставшихся n^2-n слагаемых $i\neq j$. Поэтому

$$M\left(\widehat{m}_X^2\right)=\frac{1}{n^2}\left(n\left(D_X+m_X^2\right)+\left(n^2-n\right)m_X^2\right)=\frac{1}{n}D_X+m_X^2.$$

В результате окончательно получаем

$$M\left(\widehat{D}_X\right)=\left(1-\frac{1}{n}\right)D_X,$$

т.е. оценка $\widehat{D}_X$ является *смещенной*. Чтобы получить несмещенную оценку дисперсии, переходят к скорректированному определению

$$\widehat{D}_X=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\widehat{m}_X\right)^2. \qquad (1.8)$$

Эту оценку по аналогии с предыдущим представляем как

$$\widehat{D}_X=\frac{1}{n-1}\sum_{i=1}^{n}x_i^2-\frac{n}{n-1}\widehat{m}_X^2$$

и, используя уже известные результаты, получаем

$$M\left(\widehat{D}_X\right)=\frac{n}{n-1}\left(D_X+m_X^2\right)-\frac{1}{\left(n-1\right)n}\left(n\left(D_X+m_X^2\right)+\left(n^2-n\right)m_X^2\right)=D_X,$$

т.е. оценка (1.8) оказывается *несмещенной*. Дополнительно доказывается (например, [15]), что оценка $\widehat{m}_X$ является состоятельной и эффективной, а оценка (1.8) – состоятельной.

При поиске выборочных оценок каких-либо неизвестных параметров большое значение приобретает понятие *доверительного интервала* оцениваемого параметра. Напомним существо этого термина.

Пусть найдена оценка $\widehat{a}$ параметра a. Истинное значение параметра нам не известно (иначе было бы не нужно искать его оценку). Но какие-то представления об истинном значении параметра мы хотим иметь. С этой целью и вводится термин “доверительный интервал”.

Определение 1.9. Интервал (c, d) называют доверительным для неизвестного параметра a, если выполняется условие $P(c < a < d) = 1 - \alpha$, означающее, что с вероятностью $1-\alpha$ интервал (c, d) "накрывает" истинное значение параметра a.

Величину $1-\alpha$ называют *доверительной вероятностью, α - уровнем значимости*. Доверительную вероятность можно обозначить любым другим символом, но мы здесь следуем "интернациональной" традиции, связывающей доверительную вероятность с уровнем значимости. Этими величинами задаются при построении доверительного интервала, сводя соответствующую процедуру к вычислению границ интервала *c, d*. Универсального рецепта для этих вычислений не существует и в каждом частном случае используется индивидуальный прием. Рассмотрим эти приемы при построении доверительных интервалов для параметров m_X, D_X гауссовой генеральной совокупности.

1.3. Доверительный интервал для m_X при известной дисперсии D_X

Пусть в составе функции (1.6) неизвестно только математическое ожидание m_X и поэтому получена его выборочная оценка $\widehat{m}_X$ из (1.7). Рассмотрим ошибку оценивания

$$\eta_m = \widehat{m}_X - m_X = \frac{1}{n}\sum_{i=1}^{n} x_i - m_X .$$

Легко видеть, что эта ошибка, будучи линейной комбинацией гауссовых величин, является гауссовой с нулевым математическим ожиданием и дисперсией $\frac{D_X}{n}$, т.е. $\eta_m \sim N\left(0, \frac{1}{n} D_X\right)$. Но тогда величина

$\gamma = \frac{\eta_m \sqrt{n}}{\sqrt{D_X}} \sim N(0,1)$ представляет собой стандартную гауссову величину. Обозначим ее плотность вероятностей символом $f(\gamma)$ и, учитывая симметричность графика этой функции относительно оси ординат, найдем такой интервал $(-b, b)$, на котором выполняется условие $P(\gamma \in (-b, b)) = 1 - \alpha$, где доверительная вероятность $1-\alpha$ выбрана. Выразив эту вероятность через плотность $f(\gamma)$, запишем

$$\int_{-b}^{b} f(\gamma)d\gamma = 1-\alpha \rightarrow \int_{-b}^{b} f(\gamma)d\gamma = \int_{-\infty}^{\infty} f(\gamma)d\gamma - \alpha \rightarrow$$
$$\int_{-\infty}^{-b} f(\gamma)d\gamma + \int_{b}^{\infty} f(\gamma)d\gamma = \alpha \rightarrow \int_{-\infty}^{-b} f(\gamma)d\gamma = \frac{\alpha}{2}. \qquad (1.9)$$

Финальная часть этих операций имеет следующий смысл

$$P\left(\gamma \in \left(-\infty,-b\right)\right) = \frac{\alpha}{2} \Leftrightarrow P\left(\gamma < -b\right) = \frac{\alpha}{2}.$$

Но в соответствии с определением последнее выражение означает, что величина $-b$ представляет собой квантиль $u_{\alpha/2}$ порядка $\frac{\alpha}{2}$ стандартного гауссова распределения. Следовательно, с вероятностью $1-\alpha$ выполняется неравенство $u_{\alpha/2} < \gamma < -u_{\alpha/2}$ или же с учетом определения γ $u_{\alpha/2}\sqrt{\frac{D_X}{n}} < \bar{m}_X - m_X < -u_{\alpha/2}\sqrt{\frac{D_X}{n}}$, откуда окончательно следует искомый доверительный интервал для математического ожидания

$$\bar{m}_X + u_{\alpha/2}\sqrt{\frac{D_X}{n}} < m_X < \bar{m}_X - u_{\alpha/2}\sqrt{\frac{D_X}{n}}. \qquad (1.10)$$

Если заключительную часть операций (1.9) отобразить равноправной записью $\int_{b}^{\infty} f(\gamma)d\lambda = \frac{\alpha}{2}$, имеющей смысл

$$P\left(\gamma \in \left(b,\infty\right)\right) = P\left(\gamma > b\right) = \frac{\alpha}{2},$$

то приходим к выводу, что b является $100\frac{\alpha}{2}-$процентной точкой $w_{100\alpha/2}$ стандартного гауссова распределения, и, следовательно, тот же доверительный интервал можем записать в иной редакции:

$$\bar{m}_X - w_{100\alpha/2}\sqrt{\frac{D_X}{n}} < m_X < \bar{m}_X + w_{100\alpha/2}\sqrt{\frac{D_X}{n}}. \qquad (1.11)$$

Критические точки $u_{\alpha/2}$, $w_{100\alpha/2}$ легко находятся, например, в вычислительной среде Mathcad с помощью встроенной функции

$$u_{\alpha/2} := qnorm(\frac{\alpha}{2},0,1), \quad w_{100\alpha/2} := qnorm(1-\frac{\alpha}{2},0,1). \qquad (1.12)$$

1.4. Доверительный интервал для m_X при неизвестной дисперсии D_X

Если дисперсия D_X неизвестна, то доверительный интервал для математического ожидания m_X строят аналогичным предыдущему образом, но неизвестную истинную дисперсию D_X в составе γ заменяют ее несмещенной оценкой $\widehat{D}_X$ из (1.8). В этом случае величина

$$\gamma = \frac{\sqrt{n-1}\left(\widehat{m}_X - m_x\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \widehat{m}_X\right)^2}}$$

оказывается распределенной по закону Стьюдента с *n-1* степенями свободы (например, [16]), и доверительный интервал для математического ожидания при доверительной вероятности $1-\alpha$ можно представить редакцией

$$\widehat{m}_X - w_{100\alpha/2}\sqrt{\frac{\widehat{D}_X}{n}} < m_X < \widehat{m}_X + w_{100\alpha/2}\sqrt{\frac{\widehat{D}_X}{n}}, \qquad (1.13)$$

где теперь $w_{100\alpha/2}$ – $100\dfrac{\alpha}{2}$-процентная точка распределения Стьюдента с *n-1* степенями свободы. В пакете Mathcad ее возвращает встроенная функция

$$w_{100\alpha/2} := qt(1-\frac{\alpha}{2}, n-1). \qquad (1.14)$$

1.5. Доверительный интервал для дисперсии D_X

При построении доверительного интервала для истинного значения дисперсии D_X при найденных выборочных несмещенных оценках $\widehat{m}_X$ из (1.7) и $\widehat{D}_X$ из (1.8) рассматривают статистику

$\gamma = \dfrac{\widehat{D}_X}{D_X} = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(\dfrac{x_i - \widehat{m}_X}{\sqrt{D_X}}\right)^2$, в составе которой второй сомножитель справа распределен по закону χ^2 с *n-1* степенями свободы, т.е. $z = \gamma(n-1) \sim \chi^2(n-1)$. Обозначив как $f(z)$ плотность этой величины, доверительный интервал (c, d) найдем по аналогии с предыдущим:

$$\int_c^d f(z)dz = 1-\alpha \rightarrow \int_{-\infty}^{c} f(z)dz + \int_d^{\infty} f(z)dz = \alpha .$$

Распределив α поровну между обоими слагаемыми, найдем $c = u_{\alpha/2}$, $d = w_{100\alpha/2} = u_{1-\alpha/2}$, где $u_{\alpha/2}$ и $u_{1-\alpha/2}$ - квантили соответствующих порядков χ^2 распределения с $n-1$ степенями свободы. Таким образом, получаем, что с вероятностью $1-\alpha$ выполняются неравенства $c < \dfrac{(n-1)\hat{D}_X}{D_X} < d$, из которых следует доверительный интервал

$$\frac{(n-1)\hat{D}_X}{u_{1-\alpha/2}} < D_X < \frac{(n-1)\hat{D}_X}{u_{\alpha/2}}. \qquad (1.15)$$

Критическая точка $u_{\alpha/2}$ в среде Mathcad находится как

$$u_{\alpha/2} := qchisq\left(\frac{\alpha}{2}, n-1\right). \qquad (1.16)$$

1.6. Многомерные (векторные) случайные величины

Перейдем теперь к описанию вероятностных свойств $k-$мерной случайной величины $\boldsymbol{X}$, являющейся основным объектом исследования в многомерном анализе. В значительной степени используемый при этом аппарат основывается на принципах, которыми мы руководствовались при проведении скалярного анализа, но со своими неизбежными особенностями. Начнем рассмотрение с двумерного объекта.

1.6.1. Двумерные случайные величины

Пусть $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ - двумерный вектор с компонентами X_1, X_2, представляющими собой какие-либо случайные величины, описываемые средствами nn. 1.1, 1.2. Такой вектор будем называть *двумерной случайной величиной*. Пусть, далее, $-\infty < x_1 < \infty$, $-\infty < x_2 < \infty$ - некоторые произвольные числа из указанных диапазонов. Рассмотрим два события $X_1 < x_1$, $X_2 < x_2$ и вероятность их совме-

стного выполнения $P\left(X_1 < x_1 \bigcap X_2 < x_2\right)$, т.е. вероятность того, что случайная величина X_1 окажется меньше числа x_1, а X_2 - меньше x_2. Очевидно, эта вероятность является функцией x_1 и x_2.

Определение 1.10. Двумерную функцию $F_{\boldsymbol{X}}(x_1, x_2)$ переменных x_1, x_2, равную вероятности совместного выполнения событий $X_1 < x_1$, $X_2 < x_2$, т.е. $F_{\boldsymbol{X}}\left(x_1, x_2\right) = P\left(X_1 < x_1 \bigcap X_2 < x_2\right)$, называют *двумерным законом распределения вероятностей* двумерной случайной величины $\boldsymbol{X}$.

Характерные свойства этой функции таковы:

1. Функция $F_{\boldsymbol{X}}(x_1, x_2)$ неотрицательная, неубывающая и непрерывная слева по каждому аргументу x_1, x_2.
2. $F_{\boldsymbol{X}}(-\infty, x_2) = F_{\boldsymbol{X}}(x_1, -\infty) = 0$, так как в этих случаях одно из событий в определении функции оказывается невозможным.
3. $F_{\boldsymbol{X}}(\infty, x_2) = F_{X_2}(x_2)$, $F_{\boldsymbol{X}}(x_1, \infty) = F_{X_1}(x_1)$, так как в этих случаях одно из событий в определении функции оказывается достоверным; при этом образующиеся одномерные распределения принято называть *маргинальными*.
4. $F_{\boldsymbol{X}}(\infty, \infty) = 1$.
5. $F_{\boldsymbol{X}}(x_1, x_2) = F_{\boldsymbol{X}}(x_2, x_1)$.
6. $0 \le F_{\boldsymbol{X}}(x_1, x_2) \le 1$ при $\forall x_1, x_2$.

Определение 1.11. Пусть существует функция $f_{\boldsymbol{X}}(x_1, x_2)$ такая, что можно представить

$$\int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f_{\boldsymbol{X}}(x_1, x_2)dx_1dx_2 = F_{\boldsymbol{X}}(x_1, x_2)$$

или при дифференцируемой функции $F_{\boldsymbol{X}}(x_1, x_2)$

$$f_{\boldsymbol{X}}(x_1, x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{\boldsymbol{X}}(x_1, x_2).$$

Тогда функция $f_{\boldsymbol{X}}(x_1, x_2)$ называется *двумерной плотностью вероятностей* двумерной случайной величины $\boldsymbol{X}$.

Характерные свойства этой функции:

1. $f_{\boldsymbol{X}}(x_1, x_2)dx_1dx_2 = P\left(x_1 \le X_1 < x_1 + dx_1 \bigcap x_2 \le X_2 < x_2 + dx_2\right)$.
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, x_2)dx_1dx_2 = 1$ (условие *нормировки*, вытекающее из

определения (1.11)).

3. $f_{\boldsymbol{X}}(x_1, x_2) = f_{\boldsymbol{X}}(x_2, x_1)$.

4. $f_{\boldsymbol{X}}(x_1, x_2) \geq 0$ при $\forall x_1, x_2$ как производная от неубывающей функции $F_{\boldsymbol{X}}(x_1, x_2)$.

5. $\int\limits_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, x_2)dx_1 = f_{X_2}(x_2)$, $\int\limits_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, x_2)dx_2 = f_{X_1}(x_1)$.

Чтобы убедиться, например, в справедливости первого равенства из 5, рассмотрим совокупность соотношений, основанных на свойствах функции $F_{\boldsymbol{X}}(x_1, x_2)$:

$$F_{\boldsymbol{X}}(\infty, x_2) = F_{X_2}(x_2) = \int\limits_{-\infty}^{x_2} f_{X_2}(x_2)dx_2;$$

с другой стороны,

$$F_{\boldsymbol{X}}(\infty, x_2) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{x_2} f_{\boldsymbol{X}}(x_1, x_2)dx_2dx_1 = \int\limits_{-\infty}^{x_2}\left(\int\limits_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, x_2)dx_1\right)dx_2.$$

Из сопоставления правых частей этих равенств следует первое соотношение в свойстве 5. Аналогичным образом доказывается и второе соотношение.

Определение 1.12. Функция

$$f_{X_1}(x_1 \mid x_2) = \frac{f_{\boldsymbol{X}}(x_1, x_2)}{f_{X_2}(x_2)},$$

где принимается $f_{X_2}(x_2) \neq 0$, называется *условной плотностью вероятностей* случайной величины X_1 при условии, что величина X_2 приняла значение x_2. Аналогично функция

$$f_{X_2}(x_2 \mid x_1) = \frac{f_{\boldsymbol{X}}(x_1, x_2)}{f_{X_1}(x_1)}, \quad f_{X_1}(x_1) \neq 0,$$

называется условной плотностью вероятностей случайной величины X_2 при условии, что величина X_1 приняла конкретное значение x_1.

Условные плотности $f_{X_1}(x_1 \mid x_2)$ и $f_{X_2}(x_2 \mid x_1)$ компонентов двумерного случайного вектора $\boldsymbol{X}$ не следует путать с одномерными плотностями $f_{X_1}(x_1)$ и $f_{X_2}(x_2)$ этих компонентов, так как последние вычисляются для каждого компонента самостоятельно вне зависимо-

сти от ограничения на второй компонент и потому часто называются *безусловными* плотностями. В силу этого в общем случае

$$f_{X_1}(x_1 \mid x_2) \neq f_{X_1}(x_1), \quad f_{X_2}(x_2 \mid x_1) \neq f_{X_2}(x_2)$$

и имеет место представление

$$f_{\boldsymbol{X}}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2 \mid x_1) = f_{X_2}(x_2) f_{X_1}(x_1 \mid x_2). \qquad (1.17)$$

Однако не исключаются ситуации, в которых

$$f_{X_1}(x_1 \mid x_2) = f_{X_1}(x_1),\ f_{X_2}(x_2 \mid x_1) = f_{X_2}(x_2),$$
$$f_{\boldsymbol{X}}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2), \qquad (1.18)$$

т.е. когда двумерная плотность вероятностей вектора $\boldsymbol{X}$ равняется произведению безусловных плотностей вероятности его компонентов.

Определение 1.13. Случайные величины X_1, X_2 называют *независимыми*, если их совместная плотность вероятностей равняется произведению их одномерных безусловных плотностей, т.е. выполняются равенства (1.18). В таком случае говорят, что вектор $\boldsymbol{X}$ содержит независимые компоненты.

Среди многочисленных двумерных плотностей $f_{\boldsymbol{X}}(x_1, x_2)$, используемых при решении прикладных задач, наиболее часто привлекается функция вида

$$f_{\boldsymbol{X}}(x_1, x_2) = f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^2 |\boldsymbol{K}|}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{m})^T \boldsymbol{K}^{-1} (\boldsymbol{x} - \boldsymbol{m}) \right), \qquad (1.19)$$

в определении которой $\mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ - двумерный аргумент функции $f_{\boldsymbol{X}}(x_1, x_2)$, $\boldsymbol{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$ - некоторый двумерный вектор, $\boldsymbol{K} = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix}$ - некоторая симметрическая $(k_{12} = k_{21})$ неотрицательно определенная $(\forall \boldsymbol{y} \in R^2 \ \ \boldsymbol{y}^T \boldsymbol{K} \boldsymbol{y} \geq 0)$ матрица из $R^{2\times 2}$, ^{T}- здесь и далее символ транспонирования. Смысл величин $\boldsymbol{m}$, $\boldsymbol{K}$ будет раскрыт ниже. Плотность (1.19) называют *двумерной гауссовой* или *нормальной* плотностью вероятностью вектора $\boldsymbol{X}$ и в подобных случаях по аналогии с одномерным вариантом пишут $\boldsymbol{X} \sim N(\boldsymbol{m}, \boldsymbol{K})$. Ниже, после установления смысла величин $\boldsymbol{m}$ и $\boldsymbol{K}$, будет предложена другая редакция функции $f_{\boldsymbol{X}}(\boldsymbol{x})$.

1.6.2. Числовые характеристики двумерных случайных величин

Как и в одномерном случае, при решении ряда практических задач удается ограничиться сравнительно небольшим набором сведений о вероятностных свойствах двумерной случайной величины, сконцентрированных в ее числовых характеристиках. Рассмотрим некоторые, наиболее представительные из них.

Определение 1.14. Пусть $l = l_1 + l_2$, где l_1, l_2 - целые неотрицательные величины, $\boldsymbol{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = const$, $\boldsymbol{X}$ - двумерный случайный вектор. Тогда моментом *l-го* порядка вектора $\boldsymbol{X}$ относительно постоянного вектора $\boldsymbol{c}$ называют число Q_l, определенное правилом

$$Q_l = M\left(\left(X_1 - c_1\right)^{l_1}\left(X_2 - c_2\right)^{l_2}\right) = \\ = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1 - c_1)^{l_1}(x_2 - c_2)^{l_2} f_{\boldsymbol{X}}(x_1, x_2)dx_1 dx_2. \qquad (1.20)$$

Предполагается, что интегралы в этом выражении – сходящиеся.

Из формулы (1.20) вытекает ряд важных частных случаев.

Если положить $c_1 = 0, l_1 = 1, l_2 = 0$, получим $Q_l = M\left(X_1\right) = m_{X_1}$ - математическое ожидание первого компонента вектора ***X;*** при $c_2 = 0,\ l_1 = 0, l_2 = 1$ получаем математическое ожидание второго компонента $Q_l = M\left(X_2\right) = m_{X_2}$. Если в выражении (20) $c_1 = c_2 = 0$, то Q_l называют *начальным* моментом l-го порядка вектора $\boldsymbol{X}$; при $c_1 = m_{X_1}$, $c_2 = m_{X_2}$ - *центральным* моментом l-го порядка. Центральный момент второго порядка при $l_1 = 2, l_2 = 0$ возвращает дисперсию D_{X_1} первого компонента вектора ***X,*** а при $l_1 = 0, l_2 = 2$ - дисперсию D_{X_2} второго компонента. Чтобы убедиться в этом, достаточно воспользоваться свойством функции $f_{\boldsymbol{X}}(x_1, x_2)$. Таким образом, двумерная плотность позволяет вычислить все одномерные числовые характеристики компонентов вектора ***X,*** и это вполне естественно, так как из двумерной плотности однозначно следуют одномерные прототипы. Но двумерному вектору, кроме одномерных числовых характеристик, свойственны специфичные двумерные числовые характеристики, определяющие статистическую связь между компонентами вектора.

Определение 1.15. Центральный момент второго порядка k_{X_1,X_2} при $l_1 = l_2 = 1$ называют *коэффициентом ковариации* случайных величин X_1, X_2, и, таким образом,

$$k_{X_1,X_2} = M\left(\left(X_1 - m_{X_1}\right)\left(X_2 - m_{X_2}\right)\right) = \\ = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_1 - m_{X_1}\right)\left(x_2 - m_{X_2}\right) f_{\boldsymbol{X}}\left(x_1, x_2\right) dx_1 dx_2. \tag{1.21}$$

Обратим внимание на некоторые очевидные следствия, вытекающие из этого определения: $k_{X_1,X_2} = k_{X_2,X_1}$; $k_{X_1,X_1} = D_{X_1}$; $k_{X_2,X_2} = D_{X_2}$.

Утверждение 1.1. Если компоненты X_1, X_2 вектора $\boldsymbol{X}$ независимы, то $k_{X_1,X_2} = 0$.

Действительно, при независимых компонентах X_1, X_2 имеем $f_{\boldsymbol{X}}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$, и из выражения (1.21) следует

$$k_{X_1,X_2} = \int_{-\infty}^{\infty}\left(x_1 - m_{X_1}\right) f_{X_1}(x_1) dx_1 \int_{-\infty}^{\infty}\left(x_2 - m_{X_2}\right) f_{X_2}(x_2) dx_2 = 0,$$

так как каждый из интегралов, будучи средним значением центрированной случайной величины, равен нулю.

Определение1.16. Нормированный коэффициент ковариации k_{X_1,X_2} называют коэффициентом *корреляции* r_{X_1,X_2} случайных величин X_1 и X_2, и, следовательно,

$$r_{X_1,X_2} = \frac{k_{X_1,X_2}}{\sqrt{D_{X_1} D_{X_2}}} = \frac{M\left(\left(X_1 - m_{X_1}\right)\left(X_2 - m_{X_2}\right)\right)}{\sqrt{M\left(\left(X_1 - m_{X_1}\right)^2\right) M\left(\left(X_2 - m_{X_2}\right)^2\right)}} = \\ = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_1 - m_{X_1}\right)\left(x_2 - m_{X_2}\right) f_{\boldsymbol{X}}\left(x_1, x_2\right) dx_1 dx_2}{\sqrt{\int_{-\infty}^{\infty}\left(x_1 - m_{X_1}\right)^2 f_{X_1}(x_1) dx_1 \int_{-\infty}^{\infty}\left(x_2 - m_{X_2}\right)^2 f_{\boldsymbol{X}}\left(x_2\right) dx_2}}. \tag{1.22}$$

Для этого коэффициента имеем $r_{X_1,X_2} = r_{X_2,X_1}$; $r_{X_1,X_1} = r_{X_2,X_2} = 1$. Но в отличие от коэффициента ковариации множество значений коэффициента корреляции локализовано.

Утверждение 1.2. Справедливо ограничение $-1 \le r_{X_1,X_2} \le 1$, т.е. $r_{X_1,X_2} \in [-1,1]$.

Для доказательства утверждения рассмотрим неравенство $M\left(\left(\frac{X_1 - m_{X_1}}{\sqrt{D_{X_1}}} \pm \frac{X_2 - m_{X_2}}{\sqrt{D_{X_2}}}\right)^2\right) \ge 0$. Раскрывая квадрат и усредняя, получаем неравенство $1 \pm 2r_{X_1,X_2} + 1 \ge 0$, из которого следует $r_{X_1,X_2} \ge -1$, $r_{X_1,X_2} \le 1$.

Коэффициент корреляции используется как мера статистической связи случайных компонентов X_1 и X_2 друг с другом. Если $r_{X_1,X_2} = 0$, говорят, что эти случайные величины *не коррелированные*. Если они не зависимые, то, как показано выше, и не коррелированные. Но из некоррелированности в общем случае независимость не следует. Слабую статистическую связь не удается зарегистрировать с помощью такого "грубого" в подобных случаях "прибора", как коэффициент корреляции. Однако есть случаи, в которых независимость и некоррелированность оказываются синонимами. В частности, такой случай возникает, когда величины X_1, X_2 совместно оказываются гауссовыми. Дополнительно заметим, что при $k_{X_1,X_2} > 0$ величины X_1, X_2 имеют склонность в среднем изменяться "синхронно", совместно возрастая или уменьшаясь, при $k_{X_1,X_2} < 0$ их поведение в среднем противоположно.

Предельные случаи $r_{X_1,X_2} = \pm 1$ соответствуют наличию между величинами X_1 и X_2 жесткой функциональной линейной связи. Действительно, положим, например, $X_2 = a + bX_1$, где a, b - какие-либо постоянные параметры. Легко обнаруживаем,что $m_{X_2} = a + bm_{X_1}$,

$D_{X_2} = M\left(\left(X_2 - m_{X_2}\right)^2\right) = b^2 D_{X_1}$ и, следовательно,

$r_{X_1,X_2} = \frac{M\left(\left(X_1 - m_{X_1}\right)\left(X_2 - m_{X_2}\right)\right)}{\sqrt{D_{X_1} D_{X_2}}} = \frac{bD_{X_1}}{\sqrt{b^2 D_{X_1}}} = sgn(b)$, т.е. при $b > 0$ имеем $r_{X_1,X_2} = 1$, при $b < 0$ получаем $r_{X_1,X_2} = -1$.

Во многих задачах удобнее вместо разрозненных характеристик

D_{X_1}, D_{X_2}, k_{X_1,X_2} вектора $\boldsymbol{X}$ использовать один обобщенный показатель, в качестве которого применяют *ковариационную матрицу* вектора $\boldsymbol{X}$.

Определение 1.17. Ковариационной матрицей $\boldsymbol{K_X}$ вектора $\boldsymbol{X}$ называют матрицу $\boldsymbol{K_X} = M\left(\left(\boldsymbol{X} - \boldsymbol{m_X}\right)\left(\boldsymbol{X} - \boldsymbol{m_X}\right)^T\right)$, где $\boldsymbol{m_X} = \begin{bmatrix} m_{X_1} \\ m_{X_2} \end{bmatrix}$ - вектор математических ожиданий.

Если учесть, что $\boldsymbol{X} - \boldsymbol{m_X} = \begin{bmatrix} X_1 - m_{X_1} \\ X_2 - m_{X_2} \end{bmatrix}$, то после перемножения и усреднения обнаружим

$$\boldsymbol{K_X} = \begin{bmatrix} D_{X_1} & k_{X_1,X_2} \\ k_{X_2,X_1} & D_{X_2} \end{bmatrix}, \qquad (1.23)$$

т.е. в этой матрице сконцентрированы основные числовые характеристики вектора $\boldsymbol{X}$. Матрица, легко видеть, является симметрической.

Утверждение 1.3. Ковариационная матрица является *неотрицательно определенной.*

Предварительно напомним, что квадратная матрица $\boldsymbol{A} \in R^{n\times n}$ называется неотрицательно определенной, если при $\forall \boldsymbol{b} \in R^n$ справедливо неравенство $\boldsymbol{b}^T\boldsymbol{Ab} \geq 0$. Если же при всех $\boldsymbol{b} \neq \boldsymbol{0}$ выполняется $\boldsymbol{b}^T\boldsymbol{Ab} > 0$, матрица называется положительно определенной.

Возвратимся к матрице (1.23). Зададимся произвольным вектором $\boldsymbol{b} \in R^2$ и рассмотрим произведение $\boldsymbol{b}^T\left(\boldsymbol{X} - \boldsymbol{m_X}\right)$. При любом векторе $\boldsymbol{b}$ имеем $M\left(\left(\boldsymbol{b}^T\left(\boldsymbol{X} - \boldsymbol{m_X}\right)\right)^2\right) \geq 0$. В другой редакции это неравенство можно записать так: $M\left(\boldsymbol{b}^T\left(\boldsymbol{X} - \boldsymbol{m_X}\right)\left(\boldsymbol{X} - \boldsymbol{m_X}\right)^T\boldsymbol{b}\right) \geq 0$. Вынося $\boldsymbol{b}$ и $\boldsymbol{b}^T$ из-под знака усреднения, окончательно получаем $\boldsymbol{b}^T\boldsymbol{K_X}\boldsymbol{b} \geq 0$.

Другой, не менее распространенной матричной характеристикой векторной случайной величины является ее *корреляционная матрица*. Для ее определения введем в рассмотрение *стандартизованный* вектор

$$\boldsymbol{Y} = \begin{bmatrix} \dfrac{X_1 - m_{X_1}}{\sqrt{D_{X_1}}} \\ \dfrac{X_2 - m_{X_2}}{\sqrt{D_{X_2}}} \end{bmatrix}. \qquad (1.24)$$

Определение 1.18. Корреляционной матрицей $\boldsymbol{R_X}$ векторной случайной величины $\boldsymbol{X}$ называют матрицу $\boldsymbol{R_X} = M\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)$.

Если в этом определении учесть (1.24) и провести усреднения, в явном виде получим

$$\boldsymbol{R_X} = \begin{bmatrix} 1 & r_{X_1,X_2} \\ r_{X_2,X_1} & 1 \end{bmatrix}, \qquad (1.25)$$

т.е. у корреляционной матрицы по главной диагонали располагаются единицы, а остальные элементы - взаимные корреляции компонентов вектора $\boldsymbol{X}$. Эта матрица, как и ковариационная, является симметрической и неотрицательно определенной.

Приступим теперь к установлению смысла параметров $\boldsymbol{m}$ и $\boldsymbol{K}$ в определении (1.19) двумерной гауссовой плотности вероятностей.

Утверждение 1.4. Параметры $\boldsymbol{m}, \boldsymbol{K}$ в составе гауссовой плотности (1.19) имеют смысл $\boldsymbol{m} = \boldsymbol{m_X}$, $\boldsymbol{K} = \boldsymbol{K_X}$.

Ограничим доказательство первым тезисом: $\boldsymbol{m} = \boldsymbol{m_X}$. Запишем

$$\begin{aligned} \boldsymbol{m_X} &= M(\boldsymbol{X}) = \int_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = \\ &= \frac{1}{\sqrt{(2\pi)^2 \mid \boldsymbol{K} \mid}} \int_{-\infty}^{\infty} \boldsymbol{x} \exp\left(-\frac{1}{2} \left(\boldsymbol{x} - \boldsymbol{m}\right)^T \boldsymbol{K}^{-1} \left(\boldsymbol{x} - \boldsymbol{m}\right) \right) d\boldsymbol{x}. \end{aligned} \qquad (1.26)$$

Здесь интеграл понимается как двойной и $d\boldsymbol{x} = dx_1 dx_2$. Матрица $\boldsymbol{K}^{-1}$, как обратная от симметрической матрицы, является симметрической. Известно (например, [17]), что для симметрической матрицы $\boldsymbol{K}^{-1}$ существует ортонормированная матрица $\boldsymbol{T}$, диагонализирующая эту матрицу, т.е. со свойствами $\boldsymbol{T}^T = \boldsymbol{T}^{-1}$, $|\boldsymbol{T}| = \pm 1$, $\boldsymbol{T}^T \boldsymbol{K}^{-1} \boldsymbol{T} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$, где λ_1, λ_2 – собственные числа матрицы $\boldsymbol{K}^{-1}$. Проведем в (1.26) замену

переменных $\boldsymbol{x}-\boldsymbol{m}=\boldsymbol{T}\boldsymbol{y}$, где $\boldsymbol{y}$ - новая переменная. Примем $d\boldsymbol{x}=|\boldsymbol{T}|\,d\boldsymbol{y}=d\boldsymbol{y}$ и получим

$$\boldsymbol{m}_{\boldsymbol{X}}=\frac{1}{2\pi\sqrt{|\boldsymbol{K}|}}\boldsymbol{m}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\lambda_1 y_1^2-\frac{1}{2}\lambda_2^2\right)dy_1dy_2+$$
$$+\frac{1}{2\pi\sqrt{|\boldsymbol{K}|}}\boldsymbol{T}\int_{-\infty}^{\infty}\begin{bmatrix}\boldsymbol{y_1}\\ \boldsymbol{y_2}\end{bmatrix}\exp\left(-\frac{1}{2}\lambda_1 y_1^2-\frac{1}{2}\lambda_2 y_2^2\right)dy_1dy_2, \qquad (1.27)$$

где интегралы понимаются как двойные, распадающиеся на последовательность одномерных интегралов по dy_1 и dy_2. При этом интеграл от каждого компонента вектора $\boldsymbol{y}$, умноженного на соответствующую экспоненту, как интеграл от нечетной функции в симметричных пределах равен нулю. Поэтому в целом второй интеграл в (1.27) обращается в нуль. Первый же интеграл является табличным [13] и принимает значение

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\lambda_1 y_1^2-\frac{1}{2}\lambda_2 y_2^2\right)dy_1dy_2=2\pi\sqrt{\lambda_1\lambda_2}.$$

Из матричного исчисления известно [18], что произведение собственных чисел матрицы равняется ее определителю $\lambda_1\lambda_2=\boldsymbol{K}$. Но тогда из (1.27) следует $\boldsymbol{m}=\boldsymbol{m}_{\boldsymbol{X}}$. Аналогичными средствами доказывается $\boldsymbol{K}=\boldsymbol{K}_{\boldsymbol{X}}$, т.е. матрица $\boldsymbol{K}$ в составе двумерной гауссовой плотности вектора $\boldsymbol{x}$ представляет собой ковариационную матрицу этого вектора. Если теперь выполнить все операции, предусмотренные выражением (1.19), двумерную гауссову плотность удобно отразить редакцией

$$f_{\boldsymbol{X}}(x_1,x_2)=$$
$$=\frac{1}{\sqrt{(2\pi)^2\left(D_{X_1}D_{X_2}-k_{X_1,X_2}^2\right)}}\exp\left(D_{X_1}\tilde{x}_2^2-2k_{X_1,X_2}\tilde{x}_1\tilde{x}_2+D_{X_2}\tilde{x}_1^2\right), \qquad (1.28)$$

где $\tilde{x}_1=x_1-m_{X_1}$, $\tilde{x}_2=x_2-m_{X_2}$ - центрированные величины, или лаконично $f_{\boldsymbol{X}}(\boldsymbol{x})=\square\left(\boldsymbol{m}_{\boldsymbol{X}},\boldsymbol{D}_{\boldsymbol{X}}\right)$.

Помимо перечисленных, для двумерных величин вводятся дополнительные характеристики, отражающие их двумерную природу.

Определение 1.19. Условными математическими ожиданиями случайных величин X_1 (X_2) при фиксированных значениях случайных величин X_2 (X_1) называют функции

$$m_{X_1}(x_2) = M(X_1 \mid X_2 = x_2) = \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1 \mid x_2)\, dx_1 , \qquad (1.29)$$

$$m_{X_2}(x_1) = M(X_2 \mid X_1 = x_1) = \int_{-\infty}^{\infty} x_2 f_{X_2}(x_2 \mid x_1) dx_2 . \qquad (1.30)$$

Первую из них часто называют *регрессией* X_1 на X_2, соответственно вторую – *регрессией* X_2 на X_1. Обратим внимание на то, что условные математические ожидания представляют собой функции в отличие от обычных безусловных математических ожиданий, являющихся числовыми величинами. Однако осуществим переход

$$m_{X_1} = M\left(m_{X_1}(X_2)\right) = \int_{-\infty}^{\infty} m_{X_1}(x_2) f_{X_2}(x_2) dx_2,$$

$$m_{X_2} = M\left(m_{X_2}(X_1)\right) = \int_{-\infty}^{\infty} m_{X_2}(x_1) f_{X_1}(x_1) dx_1 ,$$

в справедливости которого легко убедиться, подставив в эти выражения соотношения (1.29), (1.30) соответственно.

По аналогии с условными средними вводятся понятия *условных дисперсий*, которые определяются выражениями

$$D_{X_1}(x_2) = M\left(\left(X_1 - m_{X_1}(X_2)\right)^2 \mid X_2 = x_2\right) =$$

$$= \int_{-\infty}^{\infty} \left(x_1 - m_{X_1}(x_2)\right)^2 f_{X_1}(x_1 \mid x_2) dx_1,$$

$$D_{X_2}(x_1) = M\left(\left(X_2 - m_{X_2}(X_1)\right)^2 \mid X_1 = x_1\right) =$$

$$= \int_{-\infty}^{\infty} \left(x_2 - m_{X_2}(x_1)\right)^2 f_{X_2}(x_2 \mid x_1) dx_2.$$

Первое из этих выражений задает дисперсию величины X_1 при условии, что второй компонент X_2 вектора $\boldsymbol{X}$ принял значение x_2; соответственно второе выражение определяет условную дисперсию величины X_2 как функцию от x_1.

1.6.3. Выборочные оценки числовых характеристик двумерных случайных величин

При работе с векторными случайными величинами возникает та же проблема, что и при описании скалярных аналогов: невозможность аналитического вычисления числовых характеристик в связи с отсутствием в распоряжении исследователя плотности вероятностей $f_{\boldsymbol{X}}(x_1, x_2)$. Поэтому истинные характеристики, как и выше, заменяют их выборочными (эмпирическими) оценками. Предполагают, что получена n-мерная выборка из генеральной совокупности $f_{\boldsymbol{X}}(x_1, x_2)$, которую удобно представить в форме матрицы $\overline{\boldsymbol{X}}$ размерностью $2 \times n$ экспериментальных наблюдений:

$$\overline{\boldsymbol{X}} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \end{bmatrix}. \qquad (1.31)$$

Здесь $x_{i,j}$, $i = 1,2$, $j = \overline{1,n}$, – значение i-го компонента вектора $\boldsymbol{X}$ на $j-$м объекте. Тогда для несмещенных оценок математических ожиданий и дисперсий компонентов X_1, X_2 вектора $\boldsymbol{X}$ используют соотношения, основанные на представлениях (1.7), (1.8):

$$\hat{m}_{X_1} = \frac{1}{n}\sum_{j=1}^{n} x_{1,j}, \ \hat{m}_{X_2} = \frac{1}{n}\sum_{j=1}^{n} x_{2,j}, \qquad (1.32)$$

$$\hat{D}_{X_1} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{1,j} - \hat{m}_{X_1}\right)^2, \ \hat{D}_{X_2} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{2,j} - \hat{m}_{X_2}\right)^2. \qquad (1.33)$$

Оценки $\hat{k}_{X_1,X_2}$ и $\hat{r}_{X_1,X_2}$ коэффициентов ковариации и корреляции между компонентами также вычисляются по данным (1.31):

$$\hat{k}_{X_1,X_2} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{1,j} - \hat{m}_{X_1}\right)\left(x_{2,j} - \hat{m}_{X_2}\right), \qquad (1.34)$$

$$\hat{r}_{X_1,X_2} = \frac{\sum_{j=1}^{n}\left(x_{1,j} - \hat{m}_{X_1}\right)\left(x_{2,j} - \hat{m}_{X_2}\right)}{\sqrt{\sum_{j=1}^{n}\left(x_{1,j} - \hat{m}_{X_1}\right)^2 \sum_{j=1}^{n}\left(x_{2,j} - \hat{m}_{X_2}\right)^2}} \in [-1,1]. \qquad (1.35)$$

Оценки (1.32)-(1.35), найденные по экспериментальным данным (1.31), как правило, не равны нулю. Истинные же значения соответствующих этим оценкам параметров исследователю не известны и объ-

ективно могут как равняться нулю, так и отличаться от него. Поэтому крайне важным оказывается анализ *значимости* оценок. Оценку $\bar{a}$ некоторого параметра a будем называть *значимой*, если удастся установить, что истинное значение параметра не равно нулю, т.е. $a \neq 0$. Так как действительное значение a неизвестно, то оценка $\bar{a}$ при последующих исследованиях используется как некий эквивалент неизвестному a. Если же удастся доказать, что $a = 0$, то оценку $\bar{a}$ признаем *незначимой*, и использовать ее в каких-либо прикладных целях нецелесообразно.

Традиционно анализ значимости оценки $\bar{a}$ выполняют по следующей схеме. В рассмотрение вводят две гипотезы:

H_0 : $a = 0$, т.е. истинное значение параметра равно нулю;

H_1 : $a \neq 0$, т.е. истинное значение параметра не равно нулю.

В результате анализа следует принять одну из гипотез. Но как бы анализ ни осуществлялся, неизбежны ошибки. Выделяют ошибку 1-го рода и ошибку 2-го рода. Под ошибкой первого рода понимают принятие гипотезы H_1, когда в действительности $a = 0$ и объективно справедлива гипотеза H_0. Соответственно ошибка 2-го рода означает, что принимается решение $a = 0$ (гипотеза H_0), когда объективно справедливо $a \neq 0$ (гипотеза H_1). Вводятся вероятности этих ошибок:

$\alpha = P(H_1 \mid H_0)$ - вероятность ошибки первого рода,

$\beta = P(H_0 \mid H_1)$ - вероятность ошибки второго рода.

Величина $1-\alpha = P(H_0 \mid H_0)$ является условной вероятностью правильного решения при справедливости гипотезы H_0, т.е. вероятностью правильно указать, $\alpha = 0$, когда действительно $\alpha = 0$; соответственно $1-\beta = P(H_1 \mid H_1)$ - условная вероятность правильного решения при справедливости гипотезы H_1. Последующие усилия направляют на то, чтобы минимизировать (максимизировать) некоторый критерий, зависящий от вероятностей α, β. С этой целью пытаются найти такую величину γ, являющуюся функцией оценки $\bar{a}$ $(\gamma = \gamma(\bar{a}))$, для которой известны плотности вероятностей при выполнении обеих гипотез H_0 и H_1, т.е. функции $f(\gamma \mid H_0)$ и $f(\gamma \mid H_1)$. К сожалению, в настоящее время такую функцию $\gamma(\bar{a})$ подобрать не удается, но найдены функции $\gamma(\bar{a})$, для которых определены плотности $f(\gamma \mid H_0)$, соответствующие гипотезе H_0, и тогда проверке подвергают гипотезу H_0.

В общем случае $\gamma \in R = (-\infty, \infty)$, т.е. всей числовой оси. Множество R разбивают на два непересекающихся подмножества R_0 и R_1 $\left(R = R_0 \bigcup R_1,\ R_0 \bigcap R_1 = \varnothing\right)$, и подмножество R_0 выбирают из условия

$$P(\gamma \in R_0 \mid H_0) = \int_{R_0} f(\gamma \mid H_0)\, d\gamma = 1 - \alpha, \qquad (1.36)$$

при достижении которого вероятность правильной идентификации гипотезы H_0 равна выбранной величине $1-\alpha$. При этом правило различения гипотез H_0 и H_1 приобретает вид:

если $\gamma \in R_0 \Rightarrow H_0$, т.е. с вероятностью $1-\alpha$, считается правильным решение $a = 0$ и оценка $\hat{a}$ интерпретируется как *незначимая*;

если $\gamma \notin R_0 \Rightarrow H_1$, т.е. с вероятностью α ошибиться, принимается решение $a \neq 0$ и оценка $\hat{a}$ признается *значимой*.

Исследование значимости оценок (1.32) проведем самостоятельно для каждой из них. Так как анализ идентичен для обеих оценок, отдадим приоритет оценке $\hat{m}_{X_1}$. Рассматриваем гипотезы $H_0 : m_{X_1} = 0$ и $H_1 : m_{X_1} \neq 0$. Вводим статистику $\gamma_1 = \sqrt{\dfrac{n}{\hat{D}_{X_1}}}\hat{m}_{X_1}$, которая при справедливости гипотезы H_0 распределена по закону Стьюдента с $n-1$ степенями свободы. Обозначим эту плотность символом $f(\gamma_1 \mid H_0)$. Так как плотность Стьюдента симметрична относительно оси ординат, множество R_0 ищем в форме интервала $(-c, c)$, используя условие (1.36), т.е.

$$\int_{-c}^{c} f(\gamma_1 \mid H_0)\, d\gamma_1 = 1 - \alpha.$$

Как уже было показано в разделе 1.2, корень этого уравнения $c = w_{100\alpha/2}$, где $w_{100\alpha/2}$ – $100\dfrac{\alpha}{2}$ -процентная точка распределения Стьюдента с $n-1$ степенями свободы. Тогда алгоритм решения задачи таков: если $|\gamma_1| < c$, с вероятностью $1-\alpha$ правильным, как не противоречащим экспериментальным наблюдениям (1.31), признается решение $m_{X_1} = 0$ и оценка $\hat{m}_{X_1}$ оказывается незначимой; если $|\gamma_1| > c$, с вероятностью ошибиться α принимается альтернативное решение

$m_{X_1} \neq 0$, а оценка $\widehat{m}_{X_1}$ считается значимой. Точно такой же анализ проводят и относительно величины $\widehat{m}_{X_2}$.

В отдельных случаях на основании априорных сведений возникает предположение о том, что *истинное математическое ожидание* m_{X_1} *равняется некоторой известной величине* μ_1. С использованием наблюдений (1.31) требуется проверить это предположение. С этой целью формируются две гипотезы:

$$H_0 : m_{X_1} = \mu_1;$$

$$H_1 : m_{X_1} \neq \mu_1$$

и вычисляется статистика $\gamma = \left(\widehat{m}_{X_1} - \mu_1\right)\sqrt{\dfrac{n}{\widehat{D}_{X_1}}}$, о которой известно, что при выполнении гипотезы H_0 она распределена по закону Стьюдента с $n-1$ степенями свободы. Тогда проверяется гипотеза H_0 при выбранном уровне значимости α: если окажется $|\gamma| < w_{100\alpha/2}$, где $w_{100\alpha/2}$ - $100\alpha/2$-процентная точка распределения Стьюдента с $n-1$ степенями свободы, то с вероятностью $1-\alpha$ принимается гипотеза H_0 как не противоречащая имеющимся наблюдениям; в случае $|\gamma| > w_{100\alpha/2}$ предпочтение отдается альтернативе H_1 с вероятностью α при этом ошибиться.

При анализе значимости выборочного коэффициента корреляции $\widehat{r}_{X_1,X_2}$ в качестве "индикатора" используют функцию $\gamma = \dfrac{\widehat{r}_{X_1,X_2}\sqrt{n-2}}{\sqrt{1-\widehat{r}^2_{X_1,X_2}}}$, о которой известно, что в случае выполнения основной гипотезы $H_0 : r_{X_1,X_2} = 0$ величина $\gamma \square t(n-2)$, т.е. распределена по закону Стьюдента с $n-2$ степенями свободы. И эта гипотеза подвергается последующему анализу, схема проведения которого совершенно идентична предыдущему случаю. Поэтому и финал такой же: находится критическая точка $c = w_{100\alpha/2}$, где $w_{100\alpha/2}$ – $100\dfrac{\alpha}{2}$-процентная точка распределения Стьюдента с $n-2$ степенями свободы; если окажется $\left|\widehat{r}_{X_1,X_2}\right| < c$, то с вероятностью $1-\alpha$ считается справедливой гипотеза H_0 о некоррелированности компонентов X_1, X_2 вектора $\boldsymbol{X}$ (истинный коэффи-

циент корреляции $r_{X_1,X_2}=0$) и выборочный коэффициент корреляции признается незначимым; если же окажется $\left|\widehat{r}_{X_1,X_2}\right|>c$, то с вероятностью α ошибиться принимается альтернативное решение о коррелированности компонентов вектора $\boldsymbol{X}$ $\left(r_{X_1,X_2}\neq 0\right)$ и оценка $\widehat{r}_{X_1,X_2}$ классифицируется как значимая.

Для построения доверительного интервала для истинного коэффициента корреляции r_{X_1,X_2}, если выяснилось, что $r_{X_1,X_2}\neq 0$, применяют специальную статистику

$$z=-\frac{1}{2}\ln\frac{1+\widehat{r}_{X_1,X_2}}{1-\widehat{r}_{X_1,X_2}}, \qquad (1.37)$$

предложенную в свое время Р. Фишером. Достоинство функции z проявляется в том, что она уже при небольшом объеме n экспериментальной выборки оказывается гауссовой $z \square \square \left(m_z, D_z\right)$ с параметрами

$$m_z=\frac{1}{2}\ln\frac{1+r_{X_1,X_2}}{1-r_{X_1,X_2}}+\frac{r_{X_1,X_2}}{2(n-1)}, \quad D_z=\frac{1}{n-3}. \qquad (1.38)$$

Обратим внимание на то, что сама величина z является неслучайной функцией случайной величины $\widehat{r}_{X_1,X_2}$, но ее математическое ожидание является функцией неизвестной истинной корреляции r_{X_1,X_2}. Если задаться доверительной вероятностью $1-\alpha$, то можно найти соответствующую интервальную оценку величины z, удовлетворяющую традиционному условию

$$P\left(z_1\le z<z_2\right)=1-\alpha, \qquad (1.39)$$

где z_1, z_2 – неизвестные границы полуинтервала. Гауссова плотность симметрична относительно математического ожидания m_z. Поэтому границы z_1, z_2 будем искать в виде $z_1=m_z-\xi$, $z_2=m_z+\xi$, где ξ – подлежащая определению величина, обеспечивающая условие (1.39). В развернутой форме это равенство имеет вид

$$\frac{1}{\sqrt{2\pi D_z}}\int_{m_z-\xi}^{m_z+\xi}\exp\left(-\frac{1}{2}\left(z-m_z\right)^2\right)dz=1-\alpha\,.$$

Введя новую переменную $\dfrac{z-m_z}{\sqrt{D_z}}=s$, $dz=\sqrt{D_z}ds$, запишем

$$\frac{1}{\sqrt{2\pi}}\int_{-b}^{b}\exp\left(-\frac{s^2}{2}\right)ds = 1-\alpha,\quad b=\frac{\xi}{\sqrt{D_z}}.$$

Аналогичное соотношение нам уже встречалось при преобразовании (1.9). Используя те же самые операции, устанавливаем

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-b}\exp\left(-\frac{s^2}{2}\right)ds = \frac{\alpha}{2}.$$

Замечаем, что слева стоит интеграл от стандартной гауссовой плотности. Но тогда по определению $-b$ равняется квантилю порядка $\frac{\alpha}{2}$ этой плотности: $-b = u_{\alpha/2}$. Как следствие, находим $b = -\sqrt{D_z}u_{\alpha/2}$. Значит, с вероятностью $1-\alpha$ выполняются неравенства

$$m_z + \sqrt{D_z}u_{\alpha/2} \le z < m_z - \sqrt{D_z}u_{\alpha/2}.$$

Или же, поменяв местами z и m_z, запишем

$$z + \sqrt{D_z}u_{\alpha/2} < m_z \le z - \sqrt{D_z}u_{\alpha/2}.$$

Учитывая определение m_z из (1.38), раскрываем это выражение:

$$z + \sqrt{D_z}u_{\alpha/2} - \frac{r_{X_1,X_2}}{2(n-1)} < \frac{1}{2}\ln\frac{1+r_{X_1,X_2}}{1-r_{X_1,X_2}} \le z - \sqrt{D_z}u_{\alpha/2} - \frac{r_{X_1,X_2}}{2(n-1)}.$$

Найдем приближенное решение этого неравенства, заменив на его границах истинный коэффициент корреляции r_{X_1,X_2} его выборочной оценкой $\widehat{r}_{X_1,X_2}$. Обозначим

$$c = z + \sqrt{D_z}u_{\alpha/2} - \frac{\widehat{r}_{X_1,X_2}}{2(n-1)},\qquad d = z - \sqrt{D_z}u_{\alpha/2} - \frac{\widehat{r}_{X_1,X_2}}{2(n-1)}. \tag{1.40}$$

В этих обозначениях

$$c < \frac{1}{2}\ln\frac{1+r_{X_1,X_2}}{1-r_{X_1,X_2}} \le d.$$

Из левого неравенства следует

$$r_{X_1,X_2} > \frac{e^{2c}-1}{e^{2c}+1} = \frac{e^{c}-e^{-c}}{e^{c}+e^{-c}} = th(c),$$

где $th(c)$ - гиперболический тангенс c. По аналогии из правого неравенства получаем $r_{X_1,X_2} \le th(d)$. Следовательно, с вероятностью $1-\alpha$

$$th(c) < r_{X_1,X_2} \le th(d), \tag{1.41}$$

что и будет доверительным интервалом для истинного коэффициента корреляции. В вычислительной среде Mathcad гиперболический тангенс возвращает функция $th(c) := tanh(c)$.

1.6.4. *k*-мерные случайные величины

Изложенный аппарат описания двумерных случайных величин полностью распространяется на векторные случайные величины произвольной размерностью $k > 2$, но содержит и некоторые дополнительные понятия. Рассмотрим случайный вектор $\boldsymbol{X} \in R^k$.

Определение 1.20. k-мерную функцию $F_{\boldsymbol{X}}(x_1, x_2, ..., x_k)$ или лаконично $F_{\boldsymbol{X}}(\boldsymbol{x})$, $\boldsymbol{x} \in R^k$, представляющую собой вероятность совместного выполнения k событий вида $X_i < x_i, i = \overline{1,k}$, где $x_i \in (-\infty, \infty)$ – произвольные числа, называют k-мерным законом распределения вероятностей k-мерной случайной величины $\boldsymbol{X}$. Таким образом,

$$F_{\boldsymbol{X}}(x_1, x_2, ..., x_k) = F_{\boldsymbol{X}}(\boldsymbol{x}) = P\left(\bigcap_{i=1}^{k}(X_i < x_i)\right). \tag{1.42}$$

Эта функция: неотрицательна и не убывает при возрастании любого аргумента; не меняет своего значения при перестановке любых аргументов местами; обращается в нуль, если какой-либо аргумент (какие-либо аргументы) обращается в $-\infty$; если какой-либо (какие-либо) аргумент обращается в $+\infty$, размерность функции уменьшается и она представляет собой закон распределения вероятностей тех компонентов вектора $\boldsymbol{X}$, которые не обратились в ∞ (маргинальные распределения); $F_X(\infty, \infty, ..., \infty) = 1$; множество значений функции $F_{\boldsymbol{X}}(\boldsymbol{x})$ представлено отрезком $[0,1]$.

Определение 1.21. Если существует функция $f_{\boldsymbol{X}}(x_1, x_2, ..., x_k) = f_{\boldsymbol{X}}(\boldsymbol{x})$, $\boldsymbol{x} \in R^k$, такая, что

$$\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}...\int_{-\infty}^{x_k} f_{\boldsymbol{X}}(x_1, x_2, ..., x_k)dx_k dx_{k-1}...dx_2 dx_1 = \int_{-\infty}^{\boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} = F_{\boldsymbol{X}}(\boldsymbol{x}),$$

или при дифференцируемой функции $F_{\boldsymbol{X}}(\boldsymbol{x})$

$$f_{\boldsymbol{X}}(x_1, x_2, ..., x_k) = \frac{\partial^k}{\partial x_1 \partial x_2 ... \partial x_k} F_{\boldsymbol{X}}(x_1, x_2, ..., x_k) = \frac{\partial}{\partial x} F_{\boldsymbol{X}}(\boldsymbol{x}),$$

то функция $f_{\boldsymbol{X}}(\boldsymbol{x})$ называется k-мерной плотностью вероятностей случайного вектора $\boldsymbol{X}$.

Примечание: здесь и далее дифференциал dx векторной величины $\boldsymbol{x}$ понимается как произведение дифференциалов компонентов этого вектора, т.е. $dx = dx_1 dx_2 ... dx_k$, и соответствующий интеграл является k-мерным.

Распространив свойства двумерной плотности на случай $k > 2$, получим характерные особенности функции $f(\boldsymbol{x})$:

$$f_{\boldsymbol{X}}(x_1, x_2, ..., x_k) dx_1 dx_2 ... dx_k = f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = P\left(\bigcap_{i=0}^{k} (x_i \le X_i < x_i + dx_i) \right);$$

$f_{\boldsymbol{X}}(x_1, x_2, ..., x_k) \ge 0$ при $\forall x_i$, $i = \overline{1,k}$;

функция $f_{\boldsymbol{X}}(\boldsymbol{x})$ не изменяет своего значения при перестановке местами любых ее аргументов;

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, x_2, ..., x_k) dx_1 dx_2 ... dx_k = \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = 1 \text{ (нормировка);}$$

если эту функцию проинтегрировать в диапазонах $(-\infty, \infty)$ по каким-либо аргументам, получим плотность вероятностей оставшихся аргументов (маргинальная плотность);

$\int_S f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = P(X_1, X_2, ..., X_k \in S)$, где $S \subset R^k$ – некоторая область из пространства R^k.

Для k-мерной плотности в общем случае справедливо представление

$$\begin{aligned} &f_{\boldsymbol{X}}(x_1, x_2, ..., x_k) = \\ &= f_{X_1}(x_1) f_{X_2}(x_2 \mid x_1) f_{X_3}(x_3 \mid x_1, x_2) ... f_{X_k}(x_k \mid x_1, x_2, ..., x_{k-1}), \end{aligned} \qquad (1.43)$$

в котором сомножители справа представляют собой *условные* плотности вероятностей отдельных компонентов вектора $\boldsymbol{X}$. Если условные плотности оказываются равными безусловным, выражение (1.43) существенно упрощается:

$$f_{\boldsymbol{X}}(x_1, x_2, ..., x_k) = \prod_{i=1}^{k} f_{X_i}(x_i), \qquad (1.44)$$

т.е. k -мерная плотность равняется произведению одномерных плотностей. Тогда компоненты $X_i, i=\overline{1,k}$, вектора $\boldsymbol{X}$, как и в двумерном случае, называются *независимыми.*

Многомерную случайную величину, подобно двумерной, часто оказывается достаточным представить ее *числовыми характеристиками*: вектором математических ожиданий, дисперсиями компонентов, ковариационной и корреляционной матрицами. По аналогии с двумерным случаем вводятся соответствующие показатели:

$$\boldsymbol{m_X} = \left[m_{X_1} \quad m_{X_2} \quad ... \quad m_{X_k} \right]^T, \; m_{X_i} = M\left(X_i\right) = \int\limits_{-\infty}^{\infty} x_i f_{X_i}(x_i) dx_i = \int\limits_{-\infty}^{\infty} x_i f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x},$$

$$D_{X_i} = M\left(\left(X_i - m_{X_i}\right)^2\right) = \int\limits_{-\infty}^{\infty} \left(x_i - m_{X_i}\right)^2 f_{X_i}\left(x_i\right) dx_i = \int\limits_{-\infty}^{\infty} \left(x_i - m_{X_i}\right)^2 f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}, i = \overline{1,k},$$

$$\boldsymbol{K_X} = M\left(\left(\boldsymbol{X} - \boldsymbol{m_X}\right)\left(\boldsymbol{X} - \boldsymbol{m_X}\right)^T\right) = \begin{bmatrix} D_{X_1} & k_{X_1,X_2} & ... & k_{X_1,X_k} \\ k_{X_2,X_1} & D_{X_2} & ... & k_{X_2,X_k} \\ ... & ... & ... & ... \\ k_{X_k,X_1} & k_{X_k,X_2} & ... & D_{X_k} \end{bmatrix}, \qquad (1.45)$$

$$k_{X_i,X_j} = M\left(\left(X_i - m_{X_i}\right)\left(X_j - m_{X_j}\right)\right) = \int\limits_{-\infty}^{\infty} \left(x_i - m_{X_i}\right)\left(x_j - m_{X_j}\right) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x},$$

$$Y_i = \frac{X_i - m_{X_i}}{\sqrt{D_{X_i}}}, \;\; i = \overline{1,k}, \;\; \boldsymbol{Y} = \left[Y_1, Y_2, ..., Y_k\right]^T,$$

$$R_X = M\left(YY^T\right) = \begin{bmatrix} 1 & r_{X_1,X_2} & ... & r_{X_1,X_k} \\ r_{X_2,X_1} & 1 & ... & r_{X_2,X_k} \\ ... & ... & ... & ... \\ r_{X_k,X_1} & r_{X_k,X_2} & ... & 1 \end{bmatrix}, \; r_{X_i,X_j} = \frac{k_{X_i,X_j}}{\sqrt{D_{X_i} D_{X_j}}}. \qquad (1.46)$$

Эти характеристики обладают свойствами, присущими двумерному аналогу из п.1.6.1. Кроме того, для многомерного вектора $\boldsymbol{X} \in R^k$, $k > 2$, вводят дополнительные числовые характеристики в форме *частных коэффициентов корреляции* и *множественного коэффициента корреляции*.

Определение 1.22. Частным коэффициентом корреляции ρ_{X_i,X_j} компонентов X_i, X_j вектора $\boldsymbol{X}$ называют коэффициент корреляции между этими компонентами, очищенный (элиминированный) от влияния остальных компонентов вектора $\boldsymbol{X}$ (*мешающих параметров*).

По замыслу, для нахождения ρ_{X_i,X_j} устанавливают наилучшую в определенном смысле *линейную* связь величин X_i, X_j с остальными компонентами вектора $\boldsymbol{X}$ и затем из X_i, X_j удаляют (вычитают) эти линейные функции, получая соответствующие остатки. Частные корреляции находятся как обычные парные коэффициенты корреляции между этими остатками.

Реализация этого замысла приводит к следующему алгоритму вычисления частного коэффициента корреляции. Исходным материалом является корреляционная матрица $\boldsymbol{R_X}$ из (1.46). Выделим в ней элементы r_{X_i,X_j}, r_{X_i,X_i}, r_{X_j,X_j}. Заметим, что последние два из них равны единицам. Вычислим алгебраические дополнения $R_{i,j}$, $R_{i,i}$, $R_{j,j}$ этих элементов. Тогда получают выражение для частного коэффициента корреляции

$$\rho_{X_i,X_j} = \frac{-R_{i,j}}{\sqrt{R_{i,i}R_{j,j}}},\ i,\ j=\overline{1,k},\ i\neq j. \qquad (1.47)$$

Так, при $k=3$ имеем три частных коэффициента корреляции ρ_{X_1,X_2}, ρ_{X_1,X_3}, ρ_{X_2,X_3}, очищенные соответственно от X_3, X_2, X_1. Для каждого из них находим

$$R_{1,2}=-(r_{X_2,X_1}-r_{X_3,X_1}r_{X_2,X_3}),\ R_{1,3}=r_{X_2,X_1}r_{X_3,X_2}-r_{X_3,X_1},$$

$$R_{2,3}=-(r_{X_3,X_2}-r_{X_3,X_1}r_{X_1,X_2}),$$

$$R_{1,1}=1-r^2_{X_2,X_3},\ R_{2,2}=1-r^2_{X_1,X_3},\ R_{3,3}=1-r^2_{X_1,X_2}$$

и, следовательно,

$$\rho_{X_1,X_2}=\frac{r_{X_2,X_1}-r_{X_3,X_1}r_{X_2,X_3}}{\sqrt{\left(1-r^2_{X_2,X_3}\right)\left(1-r^2_{X_1,X_3}\right)}},\ \rho_{X_1,X_3}=\frac{r_{X_3X_1}-r_{X_2,X_1}r_{X_3,X_2}}{\sqrt{\left(1-r^2_{X_2,X_3}\right)\left(1-r^2_{X_1,X_2}\right)}},$$

$$\rho_{X_2,X_3}=\frac{r_{X_3,X_2}-r_{X_3,X_1}r_{X_1,X_2}}{\sqrt{\left(1-r^2_{X_1,X_3}\right)\left(1-r^2_{X_1,X_2}\right)}}.$$

Следует отметить, что частные коэффициенты корреляции, как и парные, принимают значения на отрезке $[-1,1]$. Если компоненты X_i, X_j не зависят от мешающих параметров, то частные и парные корреляции совпадают. Так, если при $k=3$ предположить, что X_1, X_2 не зависят от X_3, то легко обнаруживается $\rho_{X_1,X_2}=r_{X_1,X_2}$, $\rho_{X_1,X_3}=\rho_{X_2,X_3}=0$. В противном случае по результатам сопоставления парных и частных корреляций можно сделать определенные заключения о влиянии мешающих параметров на проводимое исследование.

Определение 1.23. Множественный коэффициент корреляции q_i, $i=\overline{1,k}$, является мерой статистической связи i-го компонента X_i вектора $\boldsymbol{X}$ со всеми остальными компонентами этого вектора.

Исходным материалом для вычисления множественного коэффициента корреляции также является корреляционная матрица $\boldsymbol{R_X}$ из (1.46), через которую следующим образом определяется q_i:

$$q_i=\sqrt{1-\frac{|\boldsymbol{R_X}|}{R_{i,i}}},\ i=\overline{1,k}, \qquad (1.48)$$

где $|\boldsymbol{R_X}|$ – определитель матрицы $\boldsymbol{R_X}$, $R_{i,i}$ – алгебраическое дополнение элемента $r_{X_i,X_i}=1$ этой матрицы. Так, например, при $k=3$ коэффициент корреляции q_1 между первым компонентом X_1 и остальными двумя X_2, X_3 вектора $\boldsymbol{X}$ оказывается равным

$$q_1=\sqrt{\frac{r_{X_1,X_2}\left(r_{X_1,X_2}-r_{X_1,X_3}r_{X_2,X_3}\right)+r_{X_1,X_3}\left(r_{X_1,X_3}-r_{X_1,X_2}r_{X_3,X_2}\right)}{1-r^2_{X_2,X_3}}}.$$

Из этого выражения следует, что если величина X_1 не коррелирована с X_2 и X_3 ($r_{X_1,X_2}=r_{X_1,X_3}=0$), то $q_1=0$. Если компоненты X_2 и X_3 не коррелированы, то множественный коэффициент корреляции равняется среднегеометрическому парных корреляций $q_1=\sqrt{r^2_{X_1,X_2}+r^2_{X_1,X_3}}$. В общем случае следует иметь в виду обязательное ограничение $q_i\in[0,1]$, $i=\overline{1,k}$.

1.6.5. Выборочные оценки числовых характеристик k -мерных случайных величин

Вычисление теоретических числовых характеристик многомерных случайных величин оказывается невозможным по тем же причинам, что и в двумерном случае: отсутствие необходимой априорной информации о вероятностных свойствах случайного вектора $\boldsymbol{X}$. Поэтому, как и выше, прибегают к экспериментальным наблюдениям за вектором $\boldsymbol{X}$ с последующей их обработкой методами, подобными двумерному случаю.

Предположим, что вектор $\boldsymbol{X}$ можно зарегистрировать (измерить) на n каких-либо объектах. Под "объектом" мы абстрактно понимаем все то, что содержит информацию о векторе $\boldsymbol{X}$. Следствием экспериментальных наблюдений является матрица $\overline{\boldsymbol{X}} \in R^{k\times n}$ размерностью k на n, аналогичная двумерному прототипу (1.31):

$$\overline{\boldsymbol{X}} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,j} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,j} & \dots & x_{2,n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i,1} & x_{i,2} & \dots & x_{i,j} & \dots & x_{i,n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{k,1} & x_{k,2} & \dots & x_{k,j} & \dots & x_{k,n} \end{bmatrix}, \qquad (1.49)$$

где $x_{i,j}$, $i=\overline{1,k}$, $j=\overline{1,n},$ – значение i -го компонента X_i вектора $\boldsymbol{X}$ на j -м объекте. Тогда по экспериментальным данным по аналогии с (1.32)-(1.35) вычисляют выборочные оценки:

$$\hat{m}_{X_i} = \frac{1}{n}\sum_{j=1}^{n} x_{i,j},\ \hat{D}_{X_i} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{i,j} - \hat{m}_{X_i}\right)^2,\ i=\overline{1,k}, \qquad (1.50)$$

$$\hat{k}_{X_i,X_j} = \frac{1}{n-1}\sum_{q=1}^{n}\left(x_{i,q} - \hat{m}_{X_i}\right)\left(x_{j,q} - \hat{m}_{X_j}\right),\ i=\overline{1,k-1},\ j=\overline{i+1,k}, \qquad (1.51)$$

$$\hat{r}_{X_i,X_j} = \frac{\hat{k}_{X_i,X_j}}{\sqrt{\hat{D}_{X_i}\hat{D}_{X_j}}} =$$

$$= \frac{\sum_{q=1}^{n}\left(x_{i,q} - \hat{m}_{X_i}\right)\left(x_{j,q} - \hat{m}_{X_j}\right)}{\sqrt{\sum_{q=1}^{n}\left(x_{i,q} - \hat{m}_{X_i}\right)^2 \sum_{q=1}^{n}\left(x_{j,q} - \hat{m}_{X_j}\right)^2}}, \quad i = \overline{1,k-1}, \quad j = \overline{i+1,k}. \qquad (1.52)$$

При формировании множеств значений индексов i, j в выражениях (1.51), (1.52) учитывается, что $\hat{k}_{X_i,X_i} = \hat{D}_{X_i}$, $\hat{r}_{X_i,X_i} = 1$, и симметричность соответствующих коэффициентов. Вычислив (1.51), (1.52), можем построить выборочные ковариационную $\hat{\boldsymbol{K}}_{\boldsymbol{X}}$ и корреляционную $\hat{\boldsymbol{R}}_{\boldsymbol{X}}$ матрицы по аналогии с (1.45), (1.46), заменив истинные характеристики их оценками. Однако при использовании современной вычислительной техники удобнее это делать в матричной форме. На основании матрицы (1.49) построим две модифицированные матрицы наблюдений - *центрированную* $\overline{\boldsymbol{Z}} = \left[z_{i,j} \right]$ и стандартизованную $\overline{\boldsymbol{Y}} = \left[y_{i,j} \right]$, $i = \overline{1,k}$, $j = \overline{1,n}$, в которых примем $z_{i,j} = x_{i,j} - \hat{m}_{X_i}$, $y_{i,j} = \frac{x_{i,j} - \hat{m}_{X_i}}{\sqrt{\hat{D}_{X_i}}}$. Тогда легко убедиться в правомочности представлений

$$\hat{\boldsymbol{K}}_{\boldsymbol{X}} = \frac{1}{n-1}\overline{\boldsymbol{Z}} \cdot \overline{\boldsymbol{Z}}^T, \qquad \hat{\boldsymbol{R}}_{\boldsymbol{X}} = \frac{1}{n-1}\overline{\boldsymbol{Y}} \cdot \overline{\boldsymbol{Y}}^T. \qquad (1.53)$$

Оценки $\hat{\rho}_{X_i,X_j}$ частных коэффициентов корреляции ищут в форме (1.47), но на основании выборочной корреляционной матрицы $\hat{\boldsymbol{R}}_{\boldsymbol{X}}$:

$$\hat{\rho}_{X_i,X_j} = \frac{-\hat{R}_{i,j}}{\sqrt{\hat{R}_{i,i}\hat{R}_{j,j}}}, \; i, j = \overline{1,k}, \; i \neq j. \qquad (1.54)$$

Выборочный множественный коэффициент корреляции $\hat{q}_i$ также ищется в соответствии с определением (1.48), но адаптированным к экспериментальным данным:

$$\hat{q}_i = \sqrt{1 - \frac{\left|\hat{\boldsymbol{R}}_{\boldsymbol{X}}\right|}{\hat{R}_{i,i}}},\ i = \overline{1,k}. \tag{1.55}$$

Анализ *значимости выборочных оценок* $\hat{m}_{X_i}$, $\hat{r}_{X_i,X_j}$ проводят по аналогии с двумерным случаем и по тем же правилам. Предположение о равенстве истинных математических ожиданий m_{X_i} известным величинам μ_i часто исследуется в обобщенной форме и соответствующую задачу интерпретируют как *сравнение вектора истинного математического ожидания* $\boldsymbol{m}_{\boldsymbol{X}}$ *со стандартом* $\boldsymbol{\mu}$. Формируются две традиционные гипотезы:

$$H_0 : \boldsymbol{m}_{\boldsymbol{X}} = \boldsymbol{\mu};$$
$$H_1 : \boldsymbol{m}_{\boldsymbol{X}} \neq \boldsymbol{\mu}.$$

Здесь $\boldsymbol{m}_{\boldsymbol{X}}$ – неизвестный вектор истинных средних, $\boldsymbol{\mu}$ – известный вектор стандартов. Для классификации гипотез вводится статистика, обобщающая скалярную величину γ и имеющая вид

$$T^2 = n\left(\hat{\boldsymbol{m}}_{\boldsymbol{X}} - \boldsymbol{\mu}\right)^T \hat{\boldsymbol{K}}_{\boldsymbol{X}}^{-1}\left(\hat{\boldsymbol{m}}_{\boldsymbol{X}} - \boldsymbol{\mu}\right), \tag{1.56}$$

где $\hat{\boldsymbol{m}}_{\boldsymbol{X}}$ и $\hat{\boldsymbol{K}}_{\boldsymbol{X}}$ - оценки вектора математических ожиданий и ковариационной матрицы соответственно. Квадратичную функцию (1.56) обычно называют *статистикой Хотеллинга*. Она обладает замечательным свойством: $\frac{n-k}{k(n-1)}T^2 \square F(k, n-k) \mid H_0$, т.е. стоящая слева величина при справедливости гипотезы H_0 распределена по закону Фишера с k степенями свободы числителя и $n-k$ степенями знаменателя. Тогда ищут условие, при котором справедлива гипотеза H_0. Схема традиционная. Задаются доверительной вероятностью $1-\alpha$ и находят квантиль $u_{1-\alpha}$ порядка $1-\alpha$ распределения $F(k, n-k)$. В среде Mathcad это осуществляется элементарно:
$u_{1-\alpha} := qF(1-\alpha, k, n-k)$. Тогда

$$\begin{aligned} &T^2 < \frac{k(n-1)}{n-k}u_{1-\alpha} \Rightarrow H_0\left(\boldsymbol{m}_{\boldsymbol{X}} = \boldsymbol{\mu}\right), P\left(H_0\right) = 1-\alpha; \\ &T^2 > \frac{k(n-1)}{n-k}u_{1-\alpha} \Rightarrow H_1\left(\boldsymbol{m}_{\boldsymbol{X}} \neq \boldsymbol{\mu}\right), P\left(H_0\right) = \alpha. \end{aligned} \tag{1.57}$$

Если ковариационная матрица $\boldsymbol{K}_{\boldsymbol{X}}$ известна, то статистика Хотеллинга корректируется:

$$T^2 = n\left(\widehat{\boldsymbol{m}}_{\boldsymbol{X}} - \boldsymbol{\mu}\right)^T \boldsymbol{K}_{\boldsymbol{X}}^{-1}\left(\widehat{\boldsymbol{m}}_{\boldsymbol{X}} - \boldsymbol{\mu}\right), \tag{1.58}$$

но $T^2 \mid H_0 \sim \chi^2(k)$, т.е. при выполнении гипотезы H_0 величина (1.58) распределена по закону χ^2 с k степенями свободы. Тогда алгоритм (1.57) приобретает вид

$$\begin{aligned} &T^2 < u_{1-\alpha} \Rightarrow H_0\left(\boldsymbol{m}_{\boldsymbol{X}} = \boldsymbol{\mu}\right), P\left(H_0\right) = 1-\alpha; \\ &T^2 > u_{1-\alpha} \Rightarrow H_1\left(\boldsymbol{m}_{\boldsymbol{X}} \neq \boldsymbol{\mu}\right), P\left(H_0\right) = \alpha, \end{aligned} \tag{1.59}$$

где теперь уже $u_{1-\alpha}$ – квантиль порядка $1-\alpha$ распределения $\chi^2(k)$. Эту величину возвращает встроенная функция $u_{1-\alpha} := qchisq\left(1-\alpha, k\right)$.

В связи с частными коэффициентами корреляции (1.54) возникают традиционные задачи: *анализ значимости выборочных частных коэффициентов корреляции и построение доверительных интервалов для истинных значений частных корреляций.* Эти задачи решаются точно так же, как это было сделано для парных коэффициентов корреляции (см. п. 1.6.2), но с одной существенной поправкой: во всех соотношениях и при формировании критических точек объем выборки n следует заменить на $n-(k-2)$, где $(k-2)$ – число мешающих параметров при переходе к частным корреляциям.

Использование *оценок* (1.55) множественных коэффициентов корреляции также влечет за собой *проверку их на значимость*. Проверяются две гипотезы: $H_0: q_i = 0$; $H_1: q_i \neq 0$, где q_i – истинное значение множественного коэффициента корреляции компонента X_i вектора $\boldsymbol{X}$ с остальными компонентами этого вектора. Вводится статистика

$$\gamma = \frac{n-k}{k-1}\,\frac{\widehat{q}_i^{\,2}}{1-\widehat{q}_i^{\,2}}, \tag{1.60}$$

о которой известно, что $\gamma \sim F\left(k-1, n-k\right) \mid H_0$, т.е. величина (1.60) при выполнении гипотезы H_0 распределена по закону Фишера с $k-1$ степенями свободы числителя и $n-k$ степенями свободы знаменателя. Проверяется гипотеза H_0. При доверительной вероятности $1-\alpha$ находится квантиль $u_{1-\alpha}$ порядка $1-\alpha$ распределения $F(k-1, n-k)$ и тогда:

$$\begin{aligned} &\gamma < u_{1-\alpha} \Rightarrow H_0(q_i = 0),\ P(H_0) = 1-\alpha; \\ &\gamma > u_{1-\alpha} \Rightarrow H_1(q_i \neq 0),\ P(H_0) = \alpha. \end{aligned}$$

Таким образом, если величина (1.60) превышает критическую точку $u_{1-\alpha}$, то с вероятностью ошибиться α принимается решение $q_i \neq 0$ и выборочный коэффициент множественной корреляции $\widehat{q}_i$ объявляется *значимым.*

1.7. Примеры корреляционного анализа в системе Mathcad

В следующих примерах рассматриваются показатели производственно-хозяйственной деятельности некоторого машиностроительного предприятия:

X_1 - производительность труда;
X_2 - удельный вес покупных изделий;
X_3 - премии и вознаграждения на одного работника;
X_4 - среднегодовая численность ППП;
X_5- среднегодовая стоимость ОПФ;
X_6- непроизводственные расходы.

Задание 1. Положив n=50, зададим матрицу экспериментальных данных

X =

	1	2	3	4	5	6	7	8	9	10
1	9.26	9.38	12.11	10.81	9.35	9.87	8.17	9.12	5.88	6.3
2	0.4	0.26	0.4	0.5	0.4	0.19	0.25	0.44	0.17	0.39
3	1.23	1.04	1.8	0.43	0.88	0.57	1.72	1.7	0.84	0.6
4	$2.601 \cdot 10^{4}$	$2.393 \cdot 10^{4}$	$2.259 \cdot 10^{4}$	$2.122 \cdot 10^{4}$	$7.394 \cdot 10^{3}$	$1.159 \cdot 10^{4}$	$2.661 \cdot 10^{4}$	$7.801 \cdot 10^{3}$	$1.159 \cdot 10^{4}$	$9.475 \cdot 10^{3}$
5	167.59	186.1	220.46	169.3	39.53	40.41	102.96	37.02	45.74	40.07
6	17.72	18.39	26.46	22.37	28.13	17.55	21.92	19.52	23.99	...

X =

	11	12	13	14	15	16	17	18
1	6.22	5.49	6.5	6.61	4.32	7.37	7.02	8.25
2	0.33	0.25	0.32	0.02	0.06	0.15	0.08	0.2
3	0.82	0.84	0.67	1.04	0.66	0.86	0.79	0.34
4	$1.081 \cdot 10^{4}$	$6.371 \cdot 10^{3}$	$2.676 \cdot 10^{4}$	$4.21 \cdot 10^{3}$	$3.557 \cdot 10^{3}$	$1.415 \cdot 10^{4}$	$9.872 \cdot 10^{3}$	$5.975 \cdot 10^{3}$
5	45.44	41.08	136.14	42.39	37.39	101.78	47.55	32.61
6	25.68	18.13	25.74	21.21	22.97	16.38	13.21	...

X =

	20	21	22	23	24	25	26	27
1	8.72	6.64	8.1	5.52	9.37	13.17	6.67	5.68
2	0.3	0.24	0.1	0.11	0.47	0.53	0.34	0.2
3	1.46	1.27	1.58	0.68	0.86	1.98	0.33	0.45
4	$9.166 \cdot 10^{3}$	$1.512 \cdot 10^{4}$	$1.143 \cdot 10^{4}$	$6.462 \cdot 10^{3}$	$2.463 \cdot 10^{4}$	$4.973 \cdot 10^{4}$	$1.147 \cdot 10^{4}$	$1.945 \cdot 10^{4}$
5	38.95	81.32	67.26	59.92	107.34	512.6	53.81	80.83
6	13.69	16.66	15.06	20.09	15.98	18.27	14.42	...

X =

	28	29	30	31	32	33	34	35
1	5.22	10.02	8.16	3.78	6.48	10.44	7.65	8.77
2	0.24	0.54	0.4	0.2	0.64	0.42	0.27	0.37
3	0.74	0.03	0.99	0.24	0.57	1.22	0.68	1
4	$1.896 \cdot 10^{4}$	$9.185 \cdot 10^{3}$	$1.748 \cdot 10^{4}$	$6.265 \cdot 10^{3}$	$8.81 \cdot 10^{3}$	$1.766 \cdot 10^{4}$	$1.034 \cdot 10^{4}$	$8.901 \cdot 10^{3}$
5	59.42	36.96	91.43	17.16	27.29	184.33	58.42	59.4
6	15.41	19.35	16.83	30.53	17.98	22.09	18.29	...

X =

	36	37	38	39	40	41	42	43
1	7	11.06	9.02	13.28	9.27	6.7	6.69	9.42
2	0.38	0.35	0.42	0.32	0.33	0.29	0.3	0.56
3	0.81	1.27	1.14	1.89	0.67	0.96	0.67	0.98
4	$8.402 \cdot 10^{3}$	$3.263 \cdot 10^{4}$	$3.116 \cdot 10^{4}$	$4.646 \cdot 10^{4}$	$1.383 \cdot 10^{4}$	$6.391 \cdot 10^{3}$	$1.111 \cdot 10^{4}$	$6.555 \cdot 10^{3}$
5	49.63	391.27	258.62	75.66	123.68	37.21	53.37	32.87
6	26.2	7.26	18.83	19.7	16.87	14.63	22.17	...

X =

	43	44	45	46	47	48	49	50
1	9.42	7.24	5.39	5.61	5.59	6.57	6.54	4.23
2	0.56	0.42	0.26	0.16	0.45	0.31	0.08	0.68
3	0.98	1.16	0.54	1.23	0.78	1.16	4.44	1.06
4	$6.555\cdot10^3$	$1.109\cdot10^4$	$9.484\cdot10^3$	$3.967\cdot10^3$	$1.528\cdot10^4$	$2.087\cdot10^4$	$1.942\cdot10^4$	$3.351\cdot10^3$
5	32.87	45.63	48.41	13.58	63.99	104.55	222.11	25.76
6	22.62	26.44	22.26	19.13	8.28	28.23	12.39	...

Задание 2. Вычислить выборочные оценки математических ожиданий наблюдаемых признаков.

Оценка математического ожидания случайной величины находится по формуле (1.7):

$$\widehat{m}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij},\ i=\overline{1,k};$$

где x_{ij} – значение i-го признака на j-м объекте ($i=\overline{1,k}$, $j=\overline{1,n}$).

i := 1.. 6 j := 1.. 50

$$M_i := \frac{1}{50}\cdot\sum_{j=1}^{50} X_{i,j} \qquad M = \begin{pmatrix} 7.764 \\ 0.314 \\ 1.025 \\ 1.503\times 10^4 \\ 94.394 \\ 19.382 \end{pmatrix}$$

Задание 3. Вычислить несмещенные оценки дисперсий наблюдаемых признаков.

Несмещенная оценка дисперсии случайной величины X_i находится по формуле (1.8).

Получаем:

$$\widehat{\sigma}_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\widehat{m}_i)^2,\ i=\overline{1,k};$$

$$D_i := \frac{1}{n-1}\sum_{j=1}^{50}(x_{ij}-M_i)^2; \qquad D = \begin{pmatrix} 4.788 \\ 0.022 \\ 0.434 \\ 1.02\times10^8 \\ 8.914\times10^3 \\ 26.529 \end{pmatrix}.$$

Задание 4. При доверительной вероятности 1-α=0.95 построить доверительные интервалы для истинных математических ожиданий признаков. Для математических ожиданий доверительные интервалы (1.13) оказываются равными

$$\widehat{m}_i - w_{100\alpha/2}\frac{\widehat{\sigma}_i}{\sqrt{n}} \le m_i \le \widehat{m}_i + w_{100\alpha/2}\frac{\widehat{\sigma}_i}{\sqrt{n}}, \quad i=\overline{1,k},$$

где $w_{100\alpha/2}$ - 100α/2-процентная точка распределения Стьюдента с n-1 степенями свободы, α - назначенный уровень значимости.

$1 - \alpha = 0.95$

$W := qt\left(1 - \frac{0.05}{2}, 50\right) = 2.009$

$M_i - W\cdot\frac{\sqrt{D_i}}{\sqrt{50}} =$

7.142
0.272
0.838
$1.216\cdot10^4$
67.575
17.919

$M_i + W\cdot\frac{\sqrt{D_i}}{\sqrt{50}} =$

8.385
0.356
1.212
$1.79\cdot10^4$
121.212
20.845

Задание 5. При том же уровне значимости построить доверительные интервалы (1.15) для истинных значений дисперсий признаков.

$$\frac{(n-1)\widehat{\sigma}_i^2}{u_{1-\alpha/2}} \le \sigma_i^2 \le \frac{(n-1)\widehat{\sigma}_i^2}{u_{\alpha/2}}, \; i=\overline{1,k},$$

u_p – p-квантиль χ^2-распределения с n-1 степенями свободы.

$U2 := gchisg\left(\frac{0.05}{2}, 49\right) = 31.555;$

$U1 := gchisg\left(1-\frac{0.05}{2}, 49\right) = 70.222.$

Для левого интервала

$\frac{49 \cdot D_i}{U1} =$

3.341
0.015
0.302
$7.114 \cdot 10^7$
$6.22 \cdot 10^3$
18.511

Для правого интервала

$\frac{49 \cdot D_i}{U2} =$

7.436
0.034
0.673
$1.583 \cdot 10^8$
$1.384 \cdot 10^4$
41.195

Задание 6. Вычислить парные коэффициенты корреляции всех признаков.

Выборочная (эмпирическая) оценка **парного** коэффициента корреляции компонент X_i и X_s вектора $\boldsymbol{X}$ вычисляется по правилу (1.52):

$$\widehat{r}_{is} = \frac{\sum_{j=1}^{n}(x_{ij} - \widehat{m}_i)(x_{sj} - \widehat{m}_s)}{\sqrt{\sum_{j=1}^{n}(x_{ij} - \widehat{m}_i)^2 \sum_{j=1}^{n}(x_{sj} - m_s)^2}},\ i, s = \overline{1,k},\ i \neq s.$$

Вычислим по данной формуле парные коэффициенты корреляции и получим корреляционную матрицу r.

В среде Mathcad находим

$s := 1..6$ $\quad$ $i := 1..6$

$$r_{i,s} := \frac{\sum_{j=1}^{50}\left[(X_{i,j} - M_i)(X_{s,j} - M_s)\right]}{\sqrt{\sum_{j=1}^{50}(X_{i,j} - M_i)^2 \cdot \sum_{j=1}^{50}(X_{s,j} - M_s)^2}} \qquad r = \begin{pmatrix} 1 & 0.353 & 0.262 & 0.638 & 0.582 & -0.059 \\ 0.353 & 1 & -0.134 & 0.198 & 0.185 & 0.023 \\ 0.262 & -0.134 & 1 & 0.379 & 0.431 & -0.214 \\ 0.638 & 0.198 & 0.379 & 1 & 0.784 & -0.105 \\ 0.582 & 0.185 & 0.431 & 0.784 & 1 & -0.191 \\ -0.059 & 0.023 & -0.214 & -0.105 & -0.191 & 1 \end{pmatrix}.$$

Элементы полученной матрицы являются парными коэффициентами корреляции соответствующих признаков.

Задание 7. Построить матрицу стандартизованных наблюдений. Стандартизованные наблюдения вычисляются по формуле:

$Z_{i,j} := \frac{X_{i,j} - m_i}{\left(\sqrt{\sigma_i}\right)}$ $z_{ij} = \frac{x_{ij} - \hat{m}_i}{\sqrt{\sigma_i^2}}$, $i = \overline{1,k}$, $j = \overline{1,n}$. Проведем все вычисления в среде Mathcad. Получим матрицу стандартизованных наблюдений:

Z =

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.684	0.739	1.986	1.392	0.725	0.963	0.186	0.62	-0.861	-0.669	-0.705	-1.039	-0.577	-0.527
2	0.579	-0.361	0.579	1.251	0.579	-0.832	-0.429	0.848	-0.966	0.512	0.109	-0.429	0.042	-1.974
3	0.311	0.022	1.176	-0.904	-0.221	-0.692	1.055	1.025	-0.282	-0.646	-0.312	-0.282	-0.54	0.022
4	1.087	0.882	0.748	0.613	-0.757	-0.341	1.147	-0.716	-0.341	-0.55	-0.418	-0.858	1.162	-1.072
5	0.776	0.971	1.335	0.793	-0.581	-0.572	0.091	-0.608	-0.515	-0.575	-0.519	-0.565	0.442	-0.551
6	-0.323	-0.193	1.374	0.58	1.698	-0.356	0.493	0.027	0.895	0.462	1.223	-0.243	1.234	...

Z =

	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	-1.574	-0.18	-0.34	0.222	0.177	0.437	-0.513	0.154	-1.025	0.734	2.471	-0.5	-0.952	-1.162
2	-1.705	-1.101	-1.571	-0.765	-0.765	-0.093	-0.496	-1.437	-1.369	1.05	1.453	0.176	-0.765	-0.496
3	-0.555	-0.251	-0.358	-1.041	0.873	0.66	0.371	0.842	-0.525	-0.251	1.45	-1.056	-0.874	-0.433
4	-1.137	-0.088	-0.511	-0.897	0.161	-0.581	$8 \cdot 10^{-3}$	-0.357	-0.849	0.95	3.436	-0.353	0.437	0.389
5	-0.604	0.078	-0.496	-0.654	0.094	-0.587	-0.138	-0.287	-0.365	0.137	4.43	-0.43	-0.144	-0.37
6	0.697	-0.583	-1.198	-0.952	-1.165	-1.105	-0.528	-0.839	0.137	-0.661	-0.216	-0.963	0.656	...

Z =

	29	30	31	32	33	34	35	36	37	38	39	40	41	42
1	1.031	0.181	-1.82	-0.587	1.223	-0.052	0.46	-0.349	1.506	0.574	2.521	0.688	-0.486	-0.491
2	1.52	0.579	-0.765	2.192	0.714	-0.294	0.378	0.445	0.243	0.714	0.042	0.109	-0.16	-0.093
3	-1.512	-0.054	-1.193	-0.692	0.296	-0.525	-0.039	-0.327	0.371	0.174	1.313	-0.54	-0.099	-0.54
4	-0.579	0.242	-0.868	-0.616	0.26	-0.465	-0.607	-0.657	1.742	1.597	3.113	-0.119	-0.856	-0.388
5	-0.608	-0.031	-0.818	-0.711	0.953	-0.381	-0.371	-0.474	3.144	1.739	-0.198	0.31	-0.606	-0.435
6	$3 \cdot 10^{-3}$	-0.495	2.164	-0.272	0.526	-0.212	1.295	1.324	-2.354	-0.107	0.062	-0.488	-0.923	...

Z =

	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1	1.506	0.574	2.521	0.688	-0.486	-0.491	0.757	-0.239	-1.085	-0.984	-0.993	-0.545	-0.559	-1.615
2	0.243	0.714	0.042	0.109	-0.16	-0.093	1.654	0.714	-0.361	-1.033	0.915	-0.026	-1.571	2.461
3	0.371	0.174	1.313	-0.54	-0.099	-0.54	-0.069	0.204	-0.737	0.311	-0.373	0.204	5.186	0.053
4	1.742	1.597	3.113	-0.119	-0.856	-0.388	-0.84	-0.391	-0.55	-1.096	0.025	0.579	0.434	-1.157
5	3.144	1.739	-0.198	0.31	-0.606	-0.435	-0.652	-0.516	-0.487	-0.856	-0.322	0.108	1.353	-0.727
6	-2.354	-0.107	0.062	-0.488	-0.923	0.541	0.629	1.37	0.559	-0.049	-2.155	1.718	-1.358	...

Задание 8. Построить корреляционную матрицу $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ по так называемым **стандартизованным** наблюдениям и сопоставить ее с результатами пункта 6.

Корреляционная матрица, вычисленная по стандартизованным наблюдениям, имеет вид:

$$Rx := \frac{1}{49} \cdot Z \cdot Z^{T} = \begin{pmatrix} 1 & 0.353 & 0.262 & 0.638 & 0.582 & -0.059 \\ 0.353 & 1 & -0.134 & 0.198 & 0.185 & 0.023 \\ 0.262 & -0.134 & 1 & 0.379 & 0.431 & -0.214 \\ 0.638 & 0.198 & 0.379 & 1 & 0.784 & -0.105 \\ 0.582 & 0.185 & 0.431 & 0.784 & 1 & -0.191 \\ -0.059 & 0.023 & -0.214 & -0.105 & -0.191 & 1 \end{pmatrix}.$$

Отметим, что корреляционная матрица Rx, вычисленная по стандартизованным наблюдениям, и матрица r , полученная в задании 6, полностью совпадают.

Задание 9. Воспользовавшись определением и выше найденной выборочной корреляционной матрицей $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$, вычислить частные коэффициенты корреляции признака X_1 с каждым из факторных признаков X_i, $i = \overline{2,6}$, и сопоставить полученные результаты со значениями соответствующих парных коэффициентов корреляции.

Формула для частных корреляций: $\widehat{\rho}_{is} = -\dfrac{\widehat{R}_{is}}{\sqrt{\widehat{R}_{ii}\widehat{R}_{ss}}}$,

где $\widehat{R}_{is}$ - алгебраическое дополнение *is*-го элемента матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$.

Вычислим в MathCad соответствующие алгебраические дополнения:

```
A := | for i ∈ 1..k                                                                                                    =
     |   for j ∈ 1..k
     |     n ← i
     |     m ← j
     |     Matr ← submatrix(Rx,i + 1,k,j + 1,k)  if i = 1 ∧ j = 1
     |     Matr ← submatrix(Rx,i + 1,k,1,k − 1)  if i = 1 ∧ j = k
     |     Matr ← submatrix(Rx,1,k − 1,j + 1,k)  if i = 1 ∧ j = 1
     |     Matr ← submatrix(Rx,1,k − 1,1,k − 1)  if i = k ∧ j = k
     |     Matr ← stack(submatrix(r,1,i − 1,j + 1,k),submatrix(r,i + 1,k,j + 1,k))  if j = 1 ∧ (i > 1 ∧ i < k)
     |     Matr ← stack(submatrix(r,1,i − 1,1,k − 1),submatrix(r,i + 1,k,1,k − 1))  if j = k ∧ (i > 1 ∧ i < k)
     |     Matr ← augment(submatrix(r,i + 1,k,1,j − 1),submatrix(r,i + 1,k,j + 1,k))  if i = 1 ∧ (j > 1 ∧ j < k)
     |     Matr ← augment(submatrix(r,1,k − 1,1,j − 1),submatrix(r,1,k − 1,j + 1,k))  if i = k ∧ (j > 1 ∧ j < k)
     |     otherwise
     |       A1 ← submatrix(Rx,1,n − 1,1,m − 1)
     |       A2 ← submatrix(Rx,n + 1,k,1,m − 1)
     |       P1 ← stack(A1,A2)
     |       A3 ← submatrix(Rx,1,n − 1,m + 1,l)
     |       A4 ← submatrix(Rx,n + 1,k,m + 1,l)
     |       P2 ← stack(A3,A4)
     |       Matr ← augment(P1,P2)
     |     A_{i,j} ← (−1)^{i+j} · |Matr|
     | A
```

A - матрица, элементами которой являются алгебраические дополнения элементов матрицы $\widehat{R}_X$.

$$A = \begin{pmatrix} 0.262 & -0.064 & -0.016 & -0.112 & -0.048 & -7.452\times 10^{-3} \\ -0.064 & 0.168 & 0.045 & 2.339\times 10^{-3} & -0.015 & -7.402\times 10^{-4} \\ -0.016 & 0.045 & 0.185 & -0.021 & -0.057 & 0.024 \\ -0.112 & 2.339\times 10^{-3} & -0.021 & 0.412 & -0.252 & -0.016 \\ -0.048 & -0.015 & -0.057 & -0.252 & 0.396 & 0.035 \\ -7.452\times 10^{-3} & -7.402\times 10^{-4} & 0.024 & -0.016 & 0.035 & 0.147 \end{pmatrix}$$

Таким образом, получаем следующие частные коэффициенты корреляции:

$i := 1..1$

$$\rho_{i,s} := \frac{-A_{i,s}}{\sqrt{A_{i,i}\cdot A_{s,s}}} \qquad s := 1..6$$

$\rho = (-1 \quad 0.304 \quad 0.073 \quad 0.341 \quad 0.148 \quad 0.038)$.

Задание 10. Частные коэффициенты корреляции проверить на значимость, для чего следует использовать статистику с соответствующими частным корреляциям изменениями.

Для исследования значимости выборочных парных корреляций используются статистики

$$\gamma_{is} = \frac{\widehat{r}_{is}\sqrt{n-2}}{\sqrt{1-\widehat{r}_{is}^2}},\ i,s=\overline{1,k},\ i \neq s$$

При проверке на значимость частных коэффициентов корреляции объем выборки ***n*** **следует заменить на** ***n-k+2.***

n1 := n − k + 2

$$\gamma := \left| \begin{array}{l} \text{for } i \in 1..1 \\ \quad \text{for } s \in 1..6 \\ \quad\quad \gamma_{i,s} \leftarrow \dfrac{\rho_{i,s}\cdot\sqrt{n1-2}}{\sqrt{1-(\rho_{i,s})^2}} \quad \text{if } i \neq s \\ \gamma \end{array} \right.$$

$\gamma = (0 \quad 2.113 \quad 0.483 \quad 2.405 \quad 0.99 \quad 0.252)$

Если окажется $|\gamma_{is}| > w_{100\alpha/2}$, где $w_{100\alpha/2}$ является $100\alpha/2$-процентной точкой распределения Стьюдента с n-2 степенями свободы, то принимается гипотеза о значимости коэффициента $\widehat{r}_{is}$, т.е. о коррелированности величин X_i и X_s.

Проверим это следующим образом: сформируем вектор, ненулевые элементы которого будут характеризовать значимые коэффициенты корреляции:

$w := qt\left(1 - \frac{\alpha}{2}, nl\right)$ $\quad w = 2.013$

z := for i ∈ 1..1
  for s ∈ 1..6
    $t_{i,s} \leftarrow \gamma_{i,s}$ if $|\gamma_{i,s}| > w$
    $t_{i,s} \leftarrow 0$ otherwise
  t

z = (0 2.113 0 2.405 0 0)

Таким образом, второй и четвертый коэффициенты являются значимыми.

Задание 11. Для значимых частных коэффициентов корреляции построить доверительные интервалы для истинной корреляции, снова предварительно проведя изменения, соответствующие частным корреляциям.

Если парный коэффициент корреляции $\widehat{r}_{is}$ оказался значимым, строится доверительный интервал для истинной корреляции r_{is}, определяемый неравенством: $th(c) < r_{is} \le th(d)$, где *th* – символ гиперболического тангенса и

$$c,d = 0.5\ln\left(\frac{1+\widehat{r}_{is}}{1-\widehat{r}_{is}}\right) \pm \frac{u_{\alpha/2}}{\sqrt{n-3}} - \frac{\widehat{r}_{is}}{2(n-1)}.$$

Величина $u_{\alpha/2}$ является $\alpha/2$ квантилью стандартного гауссовского распределения $N(0,1)$. Полуинтервал (*th(c), th(d)*], таким образом, с доверительной вероятностью $1-\alpha$ покрывает истинное значение r_{is}.

Находим квантиль стандартного гауссовского распределения

$u := qnorm\left(\frac{0.05}{2}, 0, 1\right)$

$u = -1.96$

В нашем случае значимыми являются два коэффициента корреляции. Находим для них левосторонние и правосторонние границы доверительного интервала:

$$c := \begin{vmatrix} i \leftarrow 1 \\ s \leftarrow 4 \\ c \leftarrow 0.5\ln\left(\frac{1+\rho_{i,s}}{1-\rho_{i,s}}\right) + \frac{u}{\sqrt{nl-3}} - \frac{\rho_{i,s}}{2\cdot(nl-1)} \quad \text{if } i \neq s \\ c \end{vmatrix}$$

$$d := \begin{vmatrix} i \leftarrow 1 \\ s \leftarrow 4 \\ d \leftarrow 0.5\ln\left(\frac{1+\rho_{i,s}}{1-\rho_{i,s}}\right) - \frac{u}{\sqrt{nl-3}} - \frac{\rho_{i,s}}{2\cdot(nl-1)} \quad \text{if } i \neq s \\ d \end{vmatrix}$$

$c = 0.052$ $\quad d = 0.65$

$l := \tanh(c)$ $\quad p := \tanh(d)$

$l = 0.052$ $\quad p = 0.572$

Вполне очевидно, что оба значимых коэффициента попадают в полученные интервалы.

Задание 12. Вычислить множественные коэффициенты корреляции каждого признака с остальными *k-1* признаками.

Множественные коэффициенты корреляции вычисляются по формуле (1.55):

$i := 1..6$

$$q_i := \sqrt{1 - \frac{|Rx|}{A_{i,i}}} \qquad q = \begin{pmatrix} 0.691 \\ 0.429 \\ 0.511 \\ 0.817 \\ 0.809 \\ 0.258 \end{pmatrix}$$

Задание 13. Применив статистику, проверить значимость каждого из *k* множественных коэффициентов корреляции и сделать соответствующие выводы.

Статистика в этом случае задается так:

$$F = \frac{1}{k-1}\frac{\hat{q}_i^2(n-k)}{1-\hat{q}_i^2},\ i = \overline{1,k}.$$

При справедливости гипотезы H_0: q_i=0 случайная величина F_i распределена по закону Фишера с *k-1* степенями свободы числителя и *n-k* степенями свободы знаменателя. Тогда, если при доверительной вероятности 1-α окажется $F_i > u_{1-\alpha}$, где $u_{1-\alpha}$ - (*1-*α)-квантиль указанного F-распределения, то гипотеза о незначимости множественного коэффициента корреляции $\widehat{q}_i$ отвергается в пользу альтернативы, т.е. с вероятностью α ошибиться принимается решение о значимости величины $\widehat{q}_i$. При противоположном неравенстве предпочтение отдается альтернативе о незначимости корреляции $\widehat{q}_i$ с вероятностью 1-α правильности этого решения.

Задаем число признаков, число объектов и формируем статистики (1.60) k=6, n=50:

$$F_i := \frac{1}{k-1} \cdot \frac{(q_i)^2 \cdot (n-k)}{1-(q_i)^2} \qquad F = \begin{pmatrix} 8.06 \\ 1.984 \\ 3.102 \\ 17.677 \\ 16.69 \\ 0.627 \end{pmatrix}$$

$$U := qF(1-0.05, 5, 50)$$

Сформируем вектор, по значению элементов которого можно судить о значимости множественных коэффициентов корреляции. Ненулевые элементы вектора будут характеризовать значимые коэффициенты множественной корреляции:

$$MK := \left| \begin{array}{l} \text{for } i \in 1..6 \\ \quad mk_i \leftarrow F_i \text{ if } F_i > U \\ \quad mk_i \leftarrow 0 \text{ if } F_i < U \\ mk \end{array} \right. \qquad MK = \begin{pmatrix} 8.06 \\ 0 \\ 3.102 \\ 17.677 \\ 16.69 \\ 0 \end{pmatrix}$$

Отметим, что с вероятностью α ошибиться принимается решение о значимости величин q1, q3, q4, q5.

1.7.1. Анализ однородности двух генеральных совокупностей

Две гауссовы генеральные совокупности $\boldsymbol{X} \square\square (\boldsymbol{m}_{\boldsymbol{X}}, \boldsymbol{K}_{\boldsymbol{X}})$ и $\boldsymbol{Y} \square\square (\boldsymbol{m}_{\boldsymbol{Y}}, \boldsymbol{K}_{\boldsymbol{Y}})$ принято называть *однородными* [1], если они имеют одинаковую размерность, одну и ту же природу компонентов и $\boldsymbol{m}_{\boldsymbol{X}} = \boldsymbol{m}_{\boldsymbol{Y}}$, $\boldsymbol{K}_{\boldsymbol{X}} = \boldsymbol{K}_{\boldsymbol{Y}}$. Задача анализа однородности заключается в следующем. Получены две подобные (1.49) матрицы экспериментальных наблюдений за векторами $\boldsymbol{X}$ и $\boldsymbol{Y}$ размерностями $k \times n1$ и $k \times n2$:

$$\overline{\boldsymbol{X}} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,j} & \dots & x_{1,n1} \\ x_{2,1} & x_{2,2} & \dots & x_{2,j} & \dots & x_{2,n1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i,1} & x_{i,2} & \dots & x_{i,j} & \dots & x_{i,n1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{k,1} & x_{k,2} & \dots & x_{k,j} & \dots & x_{k,n1} \end{bmatrix},$$

$$\overline{\boldsymbol{Y}} = \begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,j} & \dots & y_{1,n2} \\ y_{2,1} & y_{2,2} & \dots & y_{2,j} & \dots & y_{2,n2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ y_{i,1} & y_{i,2} & \dots & y_{i,j} & \dots & y_{i,n2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ y_{k,1} & y_{k,2} & \dots & y_{k,j} & \dots & y_{k,n2} \end{bmatrix},$$

на основании которых следует выяснить, однородны или нет генеральные совокупности $\boldsymbol{X}$ и $\boldsymbol{Y}$.

Задача решается в два этапа. На первом из них исследуется вопрос о равенстве ковариационных матриц $\boldsymbol{K}_{\boldsymbol{X}}$ и $\boldsymbol{K}_{\boldsymbol{Y}}$. Сопоставляются две традиционные гипотезы

$$H_0 : \boldsymbol{K}_{\boldsymbol{X}} = \boldsymbol{K}_{\boldsymbol{Y}};$$

$$H_1 : \boldsymbol{K}_{\boldsymbol{X}} \neq \boldsymbol{K}_{\boldsymbol{Y}}.$$

Для выбора одной из них, используя определения (1.50)-(1.53), находят выборочные оценки $\widehat{\boldsymbol{m}}_{\boldsymbol{X}}, \widehat{\boldsymbol{m}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{K}}_{\boldsymbol{X}}, \widehat{\boldsymbol{K}}_{\boldsymbol{Y}}$. Если отцентрированные матрицы $\boldsymbol{X}, \boldsymbol{Y}$ мысленно объединить “бок о бок”, получим расширенную матрицу экспериментальных данных размерностью $k \times (n1 + n2)$.

По ней подобным (1.53) образом можем найти еще одну несмещенную оценку $\widehat{\boldsymbol{S}}$ ковариационной матрицы при дополнительном предположении $\boldsymbol{K_X} = \boldsymbol{K_Y}$. Путем несложных операций она сводится к виду

$$\widehat{\boldsymbol{S}} = \frac{1}{n1+n2-1}\left((n1-1)\widehat{\boldsymbol{K}}_{\boldsymbol{X}} + (n2-1)\widehat{\boldsymbol{K}}_{\boldsymbol{Y}}\right). \qquad (1.61)$$

Теперь можем приступить к различению гипотез. С этой целью формируется величина $V = ba$, где

$$b = 1 - \left(\frac{1}{n1-1} + \frac{1}{n2-1} - \frac{1}{n1+n2-1}\right)\frac{2k^2+3k-1}{6(k+1)},$$

$$a = (n1+n2-1)\ln\left|\widehat{\boldsymbol{S}}\right| - (n1-1)\ln\left|\widehat{\boldsymbol{K}}_{\boldsymbol{X}}\right| - (n2-1)\left|\widehat{\boldsymbol{K}}_{\boldsymbol{Y}}\right|.$$

Известно [1], что статистика V при некоторых ограничениях и справедливости гипотезы H_0 распределена по закону χ^2 с $\frac{1}{2}k(k+1)$ степенями свободы, т.е. $V \square \chi^2\left(\frac{1}{2}k(k+1)\right) | H_0$. Тогда стандартным образом проверяют справедливость гипотезы H_0. Вычисляют величины b, a, V. Задаются доверительной вероятностью $1-\alpha$ и находят квантиль $u_{1-\alpha}$ порядка $1-\alpha$ распределения $\chi^2\left(\frac{1}{2}k(k+1)\right)$. Последствия таковы:

$$V < u_{1-\alpha} \Rightarrow H_0(\boldsymbol{K_X} = \boldsymbol{K_Y}), P(H_0) = 1-\alpha;$$

$$V > u_{1-\alpha} \Rightarrow H_1(\boldsymbol{K_X} \neq \boldsymbol{K_Y}),\ P(H_0) = \alpha.$$

При принятии альтернативы H_1, совокупности $\boldsymbol{X}, \boldsymbol{Y}$ признаются неоднородными и анализ прекращается. В случае принятия гипотезы H_0 первый этап анализа считается завершенным успешно и исследование продолжают. Обсуждается вопрос о равенстве математических ожиданий. Очередные две гипотезы таковы: $H_0 : \boldsymbol{m_X} = \boldsymbol{m_Y};\ H_1 : \boldsymbol{m_X} \neq \boldsymbol{m_Y}$. Вводят статистику Хотеллинга $T^2 = \frac{n1 \cdot n2}{n1+n2}\left(\widehat{\boldsymbol{m}}_{\boldsymbol{X}} - \widehat{\boldsymbol{m}}_{\boldsymbol{Y}}\right)^T \widehat{\boldsymbol{S}}^{-1}\left(\widehat{\boldsymbol{m}}_{\boldsymbol{X}} - \widehat{\boldsymbol{m}}_{\boldsymbol{Y}}\right)$, о которой известно, что при справедливости гипотезы H_0 величина $\frac{n1+n2-k-1}{k(n1+n2-1)}T^2$ распределена

по закону Фишера с k степенями свободы числителя и $n1+n2-k-1$ степенями свободы знаменателя, т.е.

$$\frac{n1+n2-k-1}{k(n1+n2-1)}T^2 \square F(k, n1+n2-k-1).$$

Последующие поступки носят уже известный характер. Задаются доверительной вероятностью $1-\alpha$, находят квантиль $u_{1-\alpha}$ порядка $1-\alpha$ распределения $F(k, n1+n2-k-1)$ и принимают решение:

$$T^2 < \frac{k(n1+n2-1)}{n1+n2-k-1}u_{1-\alpha} \Rightarrow H_0(\boldsymbol{m_X} = \boldsymbol{m_Y}),\ P(H_0) = 1-\alpha;$$

$$T^2 > \frac{k(n1+n1-1)}{n1+n2-k-1}u_{1-\alpha} \Rightarrow H_1(\boldsymbol{m_X} \neq \boldsymbol{m_Y}),\ P(H_0) = \alpha.$$

Если принимается гипотеза H_0, генеральные совокупности $\boldsymbol{X}$ и $\boldsymbol{Y}$ окончательно расцениваются как однородные. В противном случае, т.е. при предпочтении альтернативы H_1, однородность отсутствует.

1.8. Ранговая корреляция

1.8.1. Матрица подстановок

При работе со случайными величинами встречаются ситуации, в которых результатам наблюдения за ними не удается придать числовые значения (неколичественные признаки) или оказываются неизвестными их законы распределения вероятностей. В подобных случаях выявление статистической связи между признаками посредством обычного коэффициента корреляции или становится невозможным, или приводит к недопустимым искажениям. Прибегают к помощи так называемых ранговых корреляций. В их основе лежат понятия *ранга* наблюдения и *матрицы* подобия.

Ограничим рассмотрение двумерной случайной величиной $\boldsymbol{X}$ с компонентами X_1 и X_2. Пусть результаты n наблюдений за $\boldsymbol{X}$ представлены матрицей

$$\overline{\boldsymbol{X}} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & x_{2,3} & \dots & x_{2,n} \end{bmatrix}.$$

Рассмотрим первую строку (наблюдения за первым признаком). По этим данным построим вариационный ряд, т.е. неубывающую последовательность $x_{(1,)1} \leq x_{(\ )1} \leq x_{(\ )2} \leq \dots \leq x_{(\ n)}$. Если все неравенства в составе вариационного ряда оказываются строгими, то рангом наблюдения $x_{1,j}$ назовем номер этого наблюдения в составе вариационного

ряда. Если же неравенства нестрогие (вариационный ряд содержит несколько одинаковых членов $x_{1,j}$), то за ранг принимается среднее арифметическое номеров в составе вариационного ряда, относящихся к наблюдению $x_{1,j}$. Заметим, что в любом случае сумма всех рангов равняется $\frac{1}{2}n(n+1)$. Для упрощения последующих построений положим, что все неравенства в определении вариационного ряда строгие.

Далее, построим вариационный ряд для элементов второй строки матрицы $\bar{\boldsymbol{X}}$: $x_{(2,1)} < x_{(2,2)} < x_{(2,3)} < ... < x_{(2,n)}$, но им в соответствие поставим ранги $r(j),\ j = \overline{1,n},$ как функции соответствующих рангов первой строки.

Из рангов первого и второго признаков составляют матрицу $\boldsymbol{P}$, которую называют матрицей *подстановок*:

$$\boldsymbol{P} = \begin{bmatrix} 1 & 2 & 3 & 4 & ... & n \\ r(1) & r(2) & r(3) & r(4) & ... & r(n) \end{bmatrix}.$$

Чтобы процесс ее построения был более понятным, проиллюстрируем его следующим примером.

Пусть десяти кандидатам при приеме на работу были предложены два теста, по результатам которых кандидаты набрали следующие баллы (см. таблицу):

Тест	Результаты тестирования кандидатов (в баллах)									
1	32	15	16	18	20	28	21	29	23	17
2	12	24	23	21	20	9	11	10	15	16

По этим данным построим матрицу подстановок. Условимся считать результаты сдачи первого теста выборкой признака X_1, второго теста – выборкой признака X_2. На их основе построим два вариационных ряда:

X_1: 15 16 17 18 20 21 23 28 29 32

X_2: 9 10 11 12 15 16 20 21 23 24.

Каждому значению признака X_1 ставим в соответствие ранг, равный его номеру среди элементов вариационной последовательности. В результате получим последовательность рангов $1, 2, 3, 4, 5, 6, 7, 8, 9, 10,$ формирующих первую строку матрицы подстановок. Приступаем к формированию второй строки. Величине $x_{(1,1)} = 15$ в таблице исходных данных соответствует значение 24 при-

знака X_2. Но это число в вариационном ряду X_2 находится на десятой позиции, по этой причине устанавливаем $r(1)=10$. Величине $x_{(1,2)}=16$ в составе исходных данных соответствует значение 23 признака X_2, которое в вариационном ряду X_2 занимает девятую позицию, т.е. $r(2)=9$. Значению $x_{(1,3)}=17$ соответствует число 16 в исходных данных, идущее в вариационном ряду X_2 под номером шесть, и, следовательно, $r(3)=6$. Совершая подобные поступки и далее, заполняем вторую строку матрицы подстановок, которая оказывается равной

$$\boldsymbol{P}=\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 10 & 9 & 6 & 8 & 7 & 3 & 5 & 1 & 2 & 4 \end{bmatrix}. \qquad (1.62)$$

При ранговой корреляции, когда обычную парную корреляцию между признаками X_1 и X_2 установить не удается, ищут корреляционную связь между рангами этих признаков, т.е. между строками матрицы подстановок $\boldsymbol{P}$. Если эта матрица имеет вид

$$\boldsymbol{P}=\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{bmatrix},$$

то считается, что между признаками существует жесткая положительная связь. При

$$\boldsymbol{P}=\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 10 & 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{bmatrix}$$

эта связь расценивается как жесткая отрицательная. Промежуточные ситуации формируют иные соотношения сил статистической связи признаков.

1.8.2. Коэффициент ранговой корреляции Кендалла

В основу определения коэффициента ранговой корреляции Кендалла положены особенности второй строки матрицы подстановок $\boldsymbol{P}$. Пусть рассматриваются два элемента этой строки с рангами $r(i)$ и $r(j)$, причем $i<j$, и оказывается $r(i)>r(j)$. Тогда говорят, что в этой строке наблюдается *инверсия.* Если же обнаруживается $r(i)<r(j)$, то этот факт интерпретируется как *порядок*. Например, для элемента $r(1)=10$ матрицы (1.62) число инверсий равняется 9, а

число порядков – 0, для элемента $r(2)=9$ число инверсий – 8, число порядков - 0, для элемента $r(3)=6$ число инверсий равняется 5, число порядков – 2. И так далее.

Определение 1.24. Выборочным коэффициентом ранговой корреляции Кендалла $\hat{r}_k$ называют отношение разности чисел порядков и инверсий во второй строке матрицы подстановок к общему числу инверсий и порядков

$$\hat{r}_k = \frac{N-Q}{N+Q},$$

где N – число порядков, Q – число инверсий.

Несложно убедиться, что в общем случае $N+Q=(n-1)+(n-2)+(n-3)+...+1=\frac{n(n-1)}{2}$. Тогда, обозначив через $S=N-Q$ разность чисел порядков и инверсий, получим наиболее распространенную форму представления коэффициента Кендалла

$$\hat{r}_k = \frac{2S}{n(n-1)}. \qquad (1.63)$$

Так, в случае (1.62) непосредственно убеждаемся: $Q=38$, $N=7$, $S=-31$ и, следовательно, учитывая $n=10$, находим $\hat{r}_k=-0.689$.

Если выразить $S=2N-\frac{n(n-1)}{2}$ или $S=\frac{n(n-1)}{2}-Q$, то $\hat{r}_k$ непосредственно связывается с числом инверсий или порядков

$$\hat{r}_k = 1-\frac{4Q}{n(n-1)} = \frac{4N}{n(n-1)}-1.$$

Из этих выражений следует: если число инверсий равно нулю, то $\hat{r}_k=1$; если же число порядков равно нулю, то $\hat{r}_k=-1$. В общем случае $\hat{r}_k \in [-1,1]$. При больших объемах выборки n и при независимых признаках X_1 и X_2 случайную величину $\hat{r}_k$ допускается считать гауссовой с плотностью вероятностей $f(\hat{r}_k)=\square(0,D_k)$, где $D_k=\frac{2(2n-5)}{9n(n-1)}$. Это обстоятельство используют при анализе значимости величины $\hat{r}_k$. Вводят две гипотезы:

$$H_0:\ r_k = 0\ (X_1,\ X_2\ независимы);$$
$$H_1:\ r_k \neq 0.$$

Проверяем гипотезу H_0. На уровне значимости α ищем такой отрезок $[-a,a]$, который удовлетворяет условию

$$P(\hat{r}_k \in [-a,a] \mid H_0) = \int_{-a}^{a} f(\hat{r}_k \mid H_0)d\hat{r}_k = 1-\alpha.$$

Из этого уравнения, что неоднократно иллюстрировалось, получаем $-a = u_{\alpha/2}$, где $u_{\alpha/2}$ – квантиль порядка $\frac{\alpha}{2}$ распределения $(0, D_k)$.

Тогда: $\hat{r}_k \in (u_{\alpha/2}, -u_{\alpha/2}) \Rightarrow H_0(r_k = 0),\ P(H_0) = 1-\alpha;$

$$|\hat{r}_k| > -u_{\alpha/2} \Rightarrow H_1(r_k \neq 0),\ P(H_0) = \alpha$$

и во втором случае $\hat{r}_k$ признается значимым.

Так, в случае (1.62) находим $D_k = 0.062$ и при $\alpha = 0.01$, получаем $u_{\alpha/2} = qnorm(\alpha/2, 0, D_k) = -0.159$. Так как $|\hat{r}_k| = 0.689 > -u_{\alpha/2}$, выборочный коэффициент корреляции Кендалла признается значимым с вероятностью ошибиться 0.01 и оба теста оказываются статистически связанными друг с другом.

1.8.3. Коэффициент ранговой корреляции Спирмена

Помимо коэффициента Кендалла в эконометрической практике поучил распространение еще один вариант ранговой корреляции - коэффициент ранговой корреляции Спирмена. В основу определения этого коэффициента положена квадратичная величина S_c, которая может интерпретироваться как мера различия между строками матрицы подстановок и находится следующим образом: $S_c = \sum_{i=1}^{n} (r(i) - i)^2$. Коэффициент Спирмена $\hat{r}_c$ с ее участием определяется так:

$$\hat{r}_c = 1 - \frac{6S_c}{n^3 - n}. \qquad (1.64)$$

Например, в случае (1.62) непосредственно вычисляем $S_c = 306$ и $\hat{r}_c = -0.855$. При некоторых ограничениях и больших n приблизительно в полтора раза $|\hat{r}_c| > |\hat{r}_k|$.

При больших объемах выборки n и при независимых признаках X_1 и X_2 случайную величину $\hat{r}_k$, как и $\hat{r}_c$, считают гауссовой с плотностью вероятностей $f(\hat{r}_c) = \square\,(0, D_c)$, где $D_c = \dfrac{1}{n-1}$. Это обстоятельство используют при анализе значимости величины $\hat{r}_c$, что осуществляется аналогичным случаю $\hat{r}_k$ образом:

$$\hat{r}_c \in (u_{\alpha/2}, -u_{\alpha/2}) \Rightarrow H_0(r_c = 0),\ P(H_0) = 1-\alpha;$$
$$|\hat{r}_c| > -u_{\alpha/2} \Rightarrow H_1(r_c \neq 0),\ P(H_0) = \alpha,$$

где $u_{\alpha/2}$ – квантиль порядка $\dfrac{\alpha}{2}$ распределения $\square\,(0, D_c)$. Как обычно, при принятии гипотезы H_1 коэффициент $\hat{r}_c$ оценивается как значимый. В случае (1.62) имеем $D_c = 0.111$, $u_{\alpha/2} = -0.286$ и $|\hat{r}_c| > -u_{\alpha/2}$, т.е. коэффициент Спирмена оказывается значимым.

Коэффициент Спирмена обладает еще одним замечательным качеством: как альтернатива формуле (1.64) он вычисляется как обычный парный коэффициент корреляции между строками матрицы подстановок $\boldsymbol{P}$, т.е. по аналогии с (1.35) или с помощью соответствующей встроенной функции в каком-либо пакете прикладных программ. Так, в среде Mathcad применительно к (1.62) получаем

$$P := \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 10 & 9 & 6 & 8 & 7 & 3 & 5 & 1 & 2 & 4 \end{pmatrix}$$

$$\mathrm{corr}\left[\left(P^{T}\right)^{\langle 0\rangle}, \left(P^{T}\right)^{\langle 1\rangle}\right] = -0.855$$

Но тогда и значимость $\hat{r}_c$ можно исследовать с использованием прежнего рецепта, т.е. индикатора $\gamma = \dfrac{\hat{r}_c\sqrt{n-2}}{\sqrt{1-\hat{r}_c^2}} \square\ t(n-2) \mid H_0(r_c = 0)$. В данном примере $\gamma = -4.663$, $qt\left(1-\dfrac{\alpha}{2}, n-2\right) = 3.355$, и так как $|\gamma| > 3.355$, то и по этому критерию делается вывод о значимости коэффициента $\hat{r}_c$. Как парный коэффициент корреляции, он удовлетворяет ограничению $\hat{r}_c \in [-1,1]$.

1.8.4. Коэффициент конкордации Кендалла

Для решения многих проблем, которые сложно или невозможно отобразить количественными показателями, привлекаются эксперты, в определенной балльной шкале выражающие свои мнения об исследуемых явлениях. Тогда возникает проблема выявления согласованности в суждениях экспертов с последующей классификацией экспертов по характеру их суждений. Для решения подобных задач применяется специальный показатель, известный как *коэффициент конкордации*.

Пусть изучаются k признаков, по поводу которых свои мнения высказывают n экспертов в форме некоторых оценок $r_{i,j}$, $i=\overline{1,k}$, $j=\overline{1,n}$, где $r_{i,j}$, таким образом, – оценка j-го эксперта, выставленная им i-му признаку. Совокупность оценок упорядочим в виде $k\times n$ матрицы $R=\left[r_{i,j}\right]$, $i=\overline{1,k}$, $j=\overline{1,n}$. Найдем суммарные оценки всех экспертов, выставленных ими по каждому признаку, т.е. величины $S_i=\sum_{j=1}^{n} r_{i,j}$, $i=\overline{1,k}$, и вычислим средний балл, приходящийся на признак $\overline{S}=\frac{1}{k}\sum_{i=1}^{k} S_i$. Далее определим квадратичную меру разброса S_W экспертных мнений относительно среднего $\overline{S}$, определенную правилом $S_W=\sum_{i=1}^{k}\left(S_i-\overline{S}\right)^2$. Тогда *коэффициент конкордации Кендалла* W определяется как

$$W=\frac{12S_W}{n^2\left(k^3-k\right)}. \qquad (1.65)$$

Обязательно ограничение $W\in[0,1]$. Если $W=0$, то поступки экспертов интерпретируют как полностью рассогласованные. При $W=1$ мнения экспертов воспринимаются как совершенно согласованные. При исследовании *значимости коэффициента конкордации* используется то обстоятельство, что при справедливости гипотезы “H_0: согласованности нет” статистика $n(k-1)W \sim \chi^2(n-1)$. Тогда процедура анализа гипотезы H_0 приобретает традиционный характер: при доверительной вероятности $1-\alpha$ находят квантиль $u_{1-\alpha}$ порядка $1-\alpha$ χ^2 распределения с $n-1$ степенями свободы и

$$W < \frac{1}{n(k-1)} u_{1-\alpha} \Rightarrow H_0(\textit{согласованности нет}),\ P(H_0) = 1-\alpha,\ W \textit{ не значим};$$

при невыполнении этого неравенства с вероятностью α ошибиться

принимается решение о согласованных действиях команды экспертов и коэффициент конкордации признается значимым.

В качестве примера, заимствованного у [1], рассмотрим поведение трех судей, оценивающих по пятибалльной системе спортивные достижения пяти атлетов. Баллы, выставленные судьями спортсменам, систематизируем в форме матрицы R размерностью 5×3 $(k=5,\ \ n=3)$:

$$R = \begin{bmatrix} 2 & 1 & 3 \\ 3 & 3 & 2 \\ 1 & 2 & 1 \\ 5 & 4 & 4 \\ 4 & 5 & 5 \end{bmatrix}.$$

Суммируя по строкам, находим баллы, выставленные судьями каждому спортсмену: $S_1 = 6,\ S_2 = 8,\ S_3 = 4,\ S_4 = 13,\ S_5 = 14$, и средний балл $\overline{S} = 9$. Далее, вычисляем квадратичный разброс $S_W = 76$ и в соответствии с (1.65) находим $W = 0.844$.

Оценим значимость этого коэффициента при доверительной вероятности $1-\alpha = 0.95$. С этой целью находим квантиль $u_{0.95}$ порядка 0.95 χ^2 распределения с $n-1=2$ степенями свободы $u_{0.95} = qchisq(0.95, 2)$, $u_{0.95} = 5.991$. Обнаруживаем $W = 0.844 > \frac{u_{1-\alpha}}{n(k-1)} = 0.499$. Следовательно, с вероятностью 0.05 ошибиться принимается решение о согласованности судейских решений, а коэффициент конкордации признается значимым. При этом по результатам сопоставления величин $S_i, i = \overline{1,5}$, первое место занимает третий спортсмен, второе – первый и т.д.

1.9. Однофакторная дисперсионная модель

Пусть существует некоторый фактор X, принимающий s уровней $X_1, X_2, ..., X_s$, но численные значения этих уровней указать нельзя. Такой фактор принято называть *качественным*. Этот фактор оказывает влияние на другую величину Y, которую в подобных ситуациях интерпретируют как *результирующий признак* и при i-м уровне фактора принимают $Y \square \square \left(m_i, D_Y\right)$. При каждом качественном уровне фактора X проводятся n_j, $j = \overline{1,s}$, измерений результирующего признака Y, следствием чего является совокупность наблюдений $y_{i,j}$, $j = \overline{1,s}$, $i = \overline{1,n_j}$, $n = \sum_{j=1}^{m} n_j$. Здесь, таким образом, $y_{i,j}$ – случайное значение результирующего признака при i-м наблюдении за ним и j-м состоянии качественного фактора (j-я группа наблюдений). Задача заключается в том, чтобы по имеющимся наблюдениям выяснить, меняются ли свойства результирующего признака в зависимости от уровней, принимаемых качественным фактором, или этого влияния нет. Решению подобного рода задач посвящен специальный раздел многомерных методов, называемый *дисперсионным анализом* (например, [15,16]). Рассмотрим существо метода на примере простейшей однофакторной модели. Прежде всего, необходимо составить математическую модель задачи. Однофакторная модель основана на предположении, что среднее значение результирующего признака в общем случае зависит от состояния качественного фактора и поэтому используется представление

$$y_{i,j} = m_j + \varepsilon_{i,j}, \; i = \overline{1,n_j}, \; j = \overline{1,s}. \qquad (1.66)$$

Здесь символ m_j использован для обозначения среднего значения результирующего признака при j-м состоянии фактора X, $\varepsilon_{i,j}$ – случайное нормально распределенное отклонение от среднего, причем $M(\varepsilon_{i,j}) = 0$, $M\left(\varepsilon_{i,j}^2\right) = D_\varepsilon \; \forall i, j$. Идея дисперсионного анализ заключается в следующем [16]. В рассмотрение вводится величина, являющаяся мерой разброса результирующего признака относительно его среднего значения и обычно называемого *вариацией*. Вариация разбивается на две части: 1) сумму квадратов отклонений каждой величины от соответствующего группового среднего и 2) сумму квадратов отклонений групповых средних от общего среднего значения. Затем по опре-

деленной методологии проводится анализ этих составляющих и выносится суждение о различии или общности свойств групповых наблюдений.

Для реализации этой идеи введем обозначения

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{s} y_{i,j}\,, \qquad \hat{m}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{i,j}\,,\ \ j=\overline{1,s},$$

из которых первое представляет собой общее выборочное среднее, а второе – групповое среднее. Тогда вариация Bap определяется как

$$Bap = \sum_{i=1}^{n_j}\sum_{j=1}^{s}\left(y_{i,j}-\hat{m}\right)^2$$

и преобразуется следующим образом:

$$\sum_{i=1}^{n_j}\sum_{j=1}^{s}\left(y_{i,j}-\hat{m}\right)^2 = \sum_{i=1}^{n_j}\sum_{j=1}^{s}\left(\left(y_{i,j}-\hat{m}_j\right)-\left(\hat{m}-\hat{m}_j\right)\right)^2.$$

Раскрыв в этом выражении квадрат, рассмотрим среднее слагаемое

$$2\sum_{i=1}^{n_j}\sum_{j=1}^{s}\left(y_{i,j}-\hat{m}_j\right)\left(\hat{m}-\hat{m}_j\right) = 2\sum_{j=1}^{s}\left(\hat{m}-\hat{m}_j\right)\sum_{i=1}^{n_j}\left(y_{i,j}-\hat{m}_j\right) =$$

$$= 2\sum_{j=1}^{s}\left(\hat{m}-\hat{m}_j\right)\left(\frac{n_j}{n_j}\sum_{i=1}^{n_j} y_{i,j} - n_j\hat{m}_j\right) = 2\sum_{j=1}^{s}\left(\hat{m}-\hat{m}_j\right)\left(n_j\hat{m}_j - n_j\hat{m}_j\right) = 0,$$

в силу чего $Bap = Q_1 + Q_2$, где

$$Q_1 = \sum_{i=1}^{n_j}\sum_{j=1}^{s}\left(y_{i,j}-\hat{m}_j\right)^2, \quad Q_2 = \sum_{i=1}^{n_j}\sum_{j=1}^{s}\left(\hat{m}-\hat{m}_j\right)^2$$

и образуют две вышеуказанные части.

Далее обсуждается гипотеза $H_0 : m_1 = m_2 = \ldots = m_s$, где m_j, $j=\overline{1,s}$, - истинные математические ожидания результирующего признака при соответствующих качественных факторах, т.е. гипотеза о том, что все качественные факторы оказывают одинаковое влияние на результирующий признак. Ищутся условия, при выполнении которых гипотеза H_0 оказывается справедливой. С этой целью формируется статистика

$$\gamma = \frac{(n-1)Q_2}{(s-1)Q_1},$$

относительно которой известно [15,16], что в случае справедливости гипотезы H_0 она распределена по закону Фишера с $s-1$ степенями свободы числителя и $n-1$ степенями свободы знаменателя, т.е.

$\gamma \square F(s-1, n-1) \mid H_0$.

Тогда, как обычно, при доверительной вероятности $1-\alpha$ находят квантиль $u_{1-\alpha}$ порядка $1-\alpha$ распределения $F(s-1, n-1)$, и $\gamma < u_{1-\alpha} \Rightarrow H_0 (m_1 = m_2 = ... = m_s)$, $P(H_0) = 1-\alpha$, т.е. при выполнении этого неравенства с вероятностью $1-\alpha$ считают справедливой гипотезу о независимости результирующего признака от состояния качественного фактора. При выполнении противоположного неравенства с вероятностью ошибиться α принимают альтернативное решение о влиянии качественного фактора на результирующий признак. В последнем случае самостоятельному исследованию может быть подвергнут вопрос о выявлении конкретных уровней качественного фактора, оказывающих влияние на результирующий признак.

2. ОСНОВЫ КОМПОНЕНТНОГО АНАЛИЗА

2.1. Постановка проблемы

Пусть изучается некоторая задача, характеризуемая k случайными признаками (показателями) $X_1, X_2, ..., X_k$ или обобщенно k-мерным вектором $\boldsymbol{X} = \left[X_1 \;\; X_2 \;\; X_3 ... \;\; X_k \right]^T$. Компоненты вектора в зависимости от предметной области исследования имеют определенный содержательный смысл. Так, например, применительно к экономической проблематике может оказаться: X_1 – производительность труда, X_2 – рентабельность, X_3 – удельный вес покупных изделий, X_4 – трудоемкость единицы продукции и т.п. В общем случае величины $X_1, X_2, ..., X_k$ имеют разные единицы измерения, разные дисперсии, коррелированы друг с другом, не упорядочены по какому-либо принципу. Вектор $\boldsymbol{X}$, будучи случайным, характеризуется, подобно (1.45), (1.46), ковариационной матрицей $\boldsymbol{K_X} = M\left(\left(\boldsymbol{X} - \boldsymbol{m_X} \right) \left(\boldsymbol{X} - \boldsymbol{m_X} \right)^T \right)$ и корреляционной матрицей

$$R_X = \begin{bmatrix} 1 & r_{X_1,X_2} & r_{X_1,X_3} & ... & r_{X_1,X_k} \\ r_{X_2,X_1} & 1 & r_{X_2,X_3} & ... & r_{X_2,X_k} \\ r_{X_3,X_1} & r_{X_3,X_2} & 1 & ... & r_{X_3,X_k} \\ ... & ... & ... & ... & ... \\ r_{X_k,X_1} & r_{X_k,X_2} & r_{X_k,X_3} & ... & 1 \end{bmatrix}.$$

Компонентный анализ предназначен для перехода от k исходных признаков $\boldsymbol{X}$ к новой системе признаков, также состоящей из k, но *других* признаков, называемых *главными компонентами.* Главные компоненты являются *гипотетическими латентными признаками* в том смысле, что непосредственному физическому измерению они не подлежат и даже не всегда удается однозначно установить их прикладную направленность. Однако главные компоненты должны обладать целым рядом новых свойств, не свойственных исходным признакам, а именно: *они должны быть не коррелированными друг с другом, центрированными, упорядоченными по величине их дисперсий, причем первый главный компонент должен иметь наибольшую дисперсию, а последний, k-й, - наименьшую.* В результате проведения компонентного анализа выявляются новые, но объективно существующие закономерности, обусловленные действием как внутренних, так и внешних причин.

Таким образом, на качественном уровне представления целью компонентного анализа является разработка для признаков $\boldsymbol{X}$ *линейной* модели вида

$$\boldsymbol{X} = \boldsymbol{Af}, \qquad (2.1)$$

где $A \in R^{k\times k}$ – подлежащая определению матрица, $\boldsymbol{f}$ – вектор главных компонентов с перечисленными выше свойствами.

2.2. Основное уравнение компонентного анализа

Основой для уточнения и решения задачи компонентного анализа является предположение о том, что исходные признаки $X_1, X_2, ..., X_k$ могут быть измерены (зарегистрированы) на n каких-либо объектах (например, на n предприятиях) и следствием измерения является матрица наблюдений $\overline{\boldsymbol{X}} \in R^{k\times n}$:

$$\overline{\boldsymbol{X}} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \dots & \dots & \dots & \dots \\ x_{k,1} & x_{k,2} & \dots & x_{k,n} \end{bmatrix}, \qquad (2.2)$$

где $x_{i,j}$ – значение i-го признака X_i на j-м объекте наблюдения. Для компонентов X_i, $i = \overline{1,k}$, вектора $\boldsymbol{X}$ можно вычислить несмещенные оценки $\hat{m}_{X_i}$ математических ожиданий и $\hat{D}_{X_i}$ дисперсий

$$\widehat{m}_{X_i} = \frac{1}{n}\sum_{j=1}^{n} x_{i,j},\ \widehat{D}_{X_i} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{i,j} - \widehat{m}_{X_i}\right)^2,\ i = \overline{1,k},$$

и построить стандартизованную матрицу наблюдений

$$\overline{\boldsymbol{Y}} = \begin{bmatrix} \dfrac{x_{1,1} - \widehat{m}_{X_1}}{\sqrt{\widehat{D}_{X_1}}} & \dfrac{x_{1,2} - \widehat{m}_{X_1}}{\sqrt{\widehat{D}_{X_1}}} & \ldots & \dfrac{x_{1,n} - \widehat{m}_{X_1}}{\sqrt{\widehat{D}_{X_1}}} \\ \dfrac{x_{2,1} - \widehat{m}_{X_2}}{\sqrt{\widehat{D}_{X_2}}} & \dfrac{x_{2,2} - \widehat{m}_{X_2}}{\sqrt{\widehat{D}_{X_2}}} & \ldots & \dfrac{x_{2,n} - \widehat{m}_{X_2}}{\sqrt{\widehat{D}_{X_2}}} \\ \ldots & \ldots & \ldots & \ldots \\ \dfrac{x_{k,1} - \widehat{m}_{X_k}}{\sqrt{\widehat{D}_{X_k}}} & \dfrac{x_{k,2} - \widehat{m}_{X_k}}{\sqrt{\widehat{D}_{X_k}}} & \ldots & \dfrac{x_{k,n} - \widehat{m}_{X_k}}{\sqrt{\widehat{D}_{X_k}}} \end{bmatrix}. \qquad (2.3)$$

По аналогии с (1.53) эту матрицу используют для построения выборочной корреляционной матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ вектора $\boldsymbol{X}$

$$\widehat{\boldsymbol{R}}_{\boldsymbol{X}} = \frac{1}{n-1}\overline{\boldsymbol{Y}}\,\overline{\boldsymbol{Y}}^T = \begin{bmatrix} 1 & \widehat{r}_{X_1,X_2} & \ldots & \widehat{r}_{X_1,X_k} \\ \widehat{r}_{X_2,X_1} & 1 & \ldots & \widehat{r}_{X_2,X_k} \\ \ldots & \ldots & \ldots & \ldots \\ \widehat{r}_{X_k,X_1} & \widehat{r}_{X_k,X_2} & \ldots & 1 \end{bmatrix}, \qquad (2.4)$$

где, в соответствии с определением,

$$\widehat{r}_{X_i,X_j} = \begin{cases} \dfrac{1}{n-1}\displaystyle\sum_{q=1}^{n} \dfrac{\left(x_{i,q} - \widehat{m}_{X_i}\right)\left(x_{j,q} - \widehat{m}_{X_j}\right)}{\sqrt{\widehat{D}_{X_i}\widehat{D}_{X_j}}},\ i, j = \overline{1,k},\ i \neq j, \\ \dfrac{1}{(n-1)\widehat{D}_{X_i}}\displaystyle\sum_{q=1}^{n}\left(x_{i,q} - \widehat{m}_{X_i}\right)^2 = 1,\ i = j. \end{cases}$$

Матрица $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$, таким образом, представляет собой эмпирически найденную оценку корреляционной матрицы исходных признаков $\boldsymbol{X}$. След $Sp\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ этой матрицы, т.е. сумма диагональных элементов, $Sp\widehat{\boldsymbol{R}}_{\boldsymbol{X}} = k$ и является суммарной дисперсией всех стандартизованных исходных признаков. Именно матрица $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ является основополагающей в компонентном анализе.

Если ориентироваться на выражение (2.1), то после регистрации вектора $\boldsymbol{X}$ на n объектах и перехода к централизованной форме представления выражение (2.1) переписывается в виде

$$\overline{\boldsymbol{Y}} = \boldsymbol{AF}\,, \qquad (2.5)$$

где $\boldsymbol{F} \in R^{k\times n}$ - матрица возможных значений главных компонентов на объектах. В развернутом виде равенство (2.5) выглядит так :

$$\begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,n} \\ y_{2,1} & y_{2,2} & \dots & y_{2,n} \\ \dots & \dots & \dots & \dots \\ y_{k,1} & y_{k,2} & \dots & y_{k,n} \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ a_{2,1} & a_{2,2} & \dots & a_{2,k} \\ \dots & \dots & \dots & \dots \\ a_{k,1} & a_{k,2} & \dots & a_{k,k} \end{bmatrix} \begin{bmatrix} f_{1,1} & f_{1,2} & \dots & f_{1,n} \\ f_{2,1} & f_{2,2} & \dots & f_{2,n} \\ \dots & \dots & \dots & \dots \\ f_{k,1} & f_{k,2} & \dots & f_{k,n} \end{bmatrix}, \qquad (2.6)$$

где $f_{i,q}$ – значение i-го главного компонента на q-м объекте. Умножив j-й столбец матрицы $\boldsymbol{F}$ на i-ю строку матрицы $\boldsymbol{A}$, получим

$$y_{i,j} = \sum_{\nu=1}^{k} a_{i,\nu} f_{\nu,j}, \;\; i = \overline{1,k}, \;\; j = \overline{1,n}. \qquad (2.7)$$

Это соотношение позволяет использовать следующую терминологию: $a_{i,\nu}$ – факторная нагрузка (вес) i-го наблюдаемого исходного признака X_i $(i = \overline{1,k})$ на ν-й главный компонент f_ν, $\nu = \overline{1,k}$.

Возвратимся теперь к равенству (2.4). С учетом (2.5) оно переписывается следующим образом:

$$\widehat{\boldsymbol{R}}_{\boldsymbol{X}} = \frac{1}{n-1}\overline{\boldsymbol{Y}}\,\overline{\boldsymbol{Y}}^{T} = \frac{1}{n-1}\boldsymbol{AFF}^{T}\boldsymbol{A}^{T}. \qquad (2.8)$$

Потребуем, чтобы главные компоненты удовлетворяли ранее сформулированным ограничениям, т.е. были центрированными, нормированными в отношении дисперсий и некоррелированными (ортогональными). Все эти условия выполняются, если от элементов матрицы $\boldsymbol{F}$ потребовать

$$\sum_{j=1}^{n} f_{i,j} = 0, \;\; \forall i = \overline{1,k} \quad \text{(центрирование);}$$

$$\frac{1}{n-1}\sum_{j=1}^{n} f_{i,j}^{2} = 1, \;\; \forall i = \overline{1,k} \quad \text{(нормировка);}$$

$$\sum_{j=1}^{n} f_{i,j} f_{\nu,j} = 0, \;\; \forall i,\nu = \overline{1,k}, \;\; i \neq \nu \qquad \text{(некоррелированность).}$$

При выполнении этих ограничений непосредственным перемножением убеждаемся, что $\frac{1}{n-1}\boldsymbol{F}\boldsymbol{F}^T = \boldsymbol{E}$, где $\boldsymbol{E}$ – единичная матрица размерностью k. И тогда выражение (2.8) существенно упрощается:

$$\widehat{\boldsymbol{R}}_{\boldsymbol{X}} = \boldsymbol{A}\boldsymbol{A}^T, \qquad (2.9)$$

связывая выборочную корреляционную матрицу $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ с матрицей нагрузок $\boldsymbol{A}$. Уравнение (2.9) является основным в компонентном анализе и создает предпосылки для вычисления матрицы нагрузок.

2.3. Вычисление матрицы нагрузок $\boldsymbol{A}$

В развернутом виде равенство (2.9) выглядит следующим образом

$$\begin{bmatrix} 1 & \widehat{r}_{X_1,X_2} & \dots & \widehat{r}_{X_1,X_k} \\ \widehat{r}_{X_2,X_1} & 1 & \dots & \widehat{r}_{X_2,X_k} \\ \dots & \dots & \dots & \dots \\ \widehat{r}_{X_k,X_1} & \widehat{r}_{X_k,X_2} & \dots & 1 \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ a_{2,1} & a_{2,2} & \dots & a_{2,k} \\ \dots & \dots & \dots & \dots \\ a_{k,1} & a_{k,2} & \dots & a_{k,k} \end{bmatrix} \begin{bmatrix} a_{1,1} & a_{2,1} & \dots & a_{k,1} \\ a_{1,2} & a_{2,2} & \dots & a_{k,2} \\ \dots & \dots & \dots & \dots \\ a_{1,k} & a_{2,k} & \dots & a_{k,k} \end{bmatrix}.$$

Матрица $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ содержит k^2 *известных* элементов, матрица $\boldsymbol{A}$ – такое же количество, но неизвестных элементов. Перемножив матрицы $\boldsymbol{A}$ и $\boldsymbol{A}^T$ и приравняв элементы матрицы $\boldsymbol{A}\boldsymbol{A}^T$ соответствующим элементам матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$, получим систему нелинейных алгебраических уравнений^

$$\sum_{j=1}^{k} a_{i,j}^2 = 1, \; i = \overline{1,k}, \qquad (2.10)$$

$$\sum_{\nu=1}^{k} a_{i,\nu} a_{j,\nu} = \widehat{r}_{i,j}, \;\; i, j = \overline{1,k}, \; i \neq j. \qquad (2.11)$$

Формально система (2.10), (2.11) состоит из k^2 уравнений и содержит такое же число неизвестных. Но однозначно найти эти неизвестные не удается, так как из-за симметричности матриц $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ и $\boldsymbol{A}\boldsymbol{A}^T$ часть уравнений в системе (2.10), (2.11) повторяется и, следовательно, число независимых уравнений оказывается меньше количества элементов в матрице $\boldsymbol{A}$. Дополнительно трудно найти решение, при котором главные компоненты обеспечивают уменьшающийся их вклад в суммарную дисперсию стандартизованных исходных признаков. Поэтому поиск матрицы $\boldsymbol{A}$ осуществляется на иных принципах.

Обратим, прежде всего, внимание на соотношения (2.11). Если их просуммировать по индексу i, то получим

$$\sum_{i=1}^{k}\sum_{j=1}^{k} a_{i,j}^2 = k, \qquad (2.12)$$

т.е. сумма квадратов элементов матрицы $\boldsymbol{A}$ равняется суммарной дисперсии k всех стандартизованных исходных признаков. Тогда логично величину $\sum_{i=1}^{k} a_{i,1}^2$ (сумму квадратов элементов первого столбца матрицы $\boldsymbol{A}$) рассматривать как вклад первого главного компонента в суммарную дисперсию, величину $\sum_{i=1}^{k} a_{i,2}^2$ – как вклад второго главного компонента и так далее, величину $\sum_{i=1}^{k} a_{i,k}^2$ – как вклад последнего k-го главного компонента. Обозначим вклады $\sum_{i=1}^{k} a_{i,j}^2 = \lambda_j$, $j = \overline{1,k}$, и будем матрицу $\boldsymbol{A}$ искать так, чтобы выполнялись неравенства $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_k$, т.е. чтобы вклад первого главного компонента был не меньше, чем второго, второго – не меньше третьего и т.д.

Матрица $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ является симметрической. С такими матрицами связана очень важная для нас теорема (например, [17]): *для симметрической матрицы* $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ *существует ортогональная матрица* $\boldsymbol{U}$, *которая диагонализирует матрицу* $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$.

Напомним, что матрицу $\boldsymbol{U}$ называют *ортогональной*, если $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{E}$, где $\boldsymbol{E}$ – единичная матрица. У ортогональной матрицы, таким образом, $\boldsymbol{U}^{-1} = \boldsymbol{U}^T$ – дополнительно определитель ортогональной матрицы $|\boldsymbol{U}| = \pm 1$, ее столбцы (строки) представляют собой ортогональные векторы (их скалярные произведения равны нулю).

Раскроем теперь в явной форме содержание теоремы: под диагонализацией матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$ понимается сведение ее к диагональной матрице $\boldsymbol{S}$, составленной из собственных чисел s_i, $i = \overline{1,k}$, матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$, по правилу

$$\boldsymbol{U}^T \hat{\boldsymbol{R}}_{\boldsymbol{X}} \boldsymbol{U} = \boldsymbol{S} = \begin{bmatrix} s_1 & 0 & ... & 0 \\ 0 & s_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & s_k \end{bmatrix}. \qquad (2.13)$$

Для построения матрицы $\boldsymbol{U}$ следует найти все собственные числа s_i, $i = \overline{1,k}$, матрицы $\hat{\boldsymbol{R}}_{\boldsymbol{X}}$, для чего следует решить алгебраическое, называемое характеристическим, уравнение k-го порядка

$$\left| \hat{\boldsymbol{R}}_{\boldsymbol{X}} - s\boldsymbol{E} \right| = 0,$$

где $|...|$ – символ определителя и $\boldsymbol{E}$ – единичная матрица;

из системы однородных алгебраических уравнений

$$\left(\hat{\boldsymbol{R}}_{\boldsymbol{X}} - s_i \boldsymbol{E} \right) \boldsymbol{C}_i = \boldsymbol{0}$$

найти соответствующие s_i ненулевые собственные векторы $\boldsymbol{C}_i$, которые при этом независимо от кратности собственного числа оказываются линейно независимыми; упорядочив эти векторы как столбцы, составить из них матрицу. Это и будет ортогональная матрица $\boldsymbol{U}$. Если дополнительно собственные векторы нормировать, получим *ортонормированную* матрицу $\boldsymbol{U}$.

Примечание. Матрица $\hat{\boldsymbol{R}}_{\boldsymbol{X}}$ является положительно определенной (точнее, неотрицательно определенной). Поэтому все ее собственные числа вещественные и неотрицательные [18]. Введем обязательное для компонентного анализа условие: перед построением матрицы $\boldsymbol{U}$, точнее перед вычислением собственных векторов, собственные числа упорядочены в направлении невозрастания, так что в (2.13) $s_1 \geq s_2 \geq ... \geq s_k$.

Рассмотрим соответствующую процедуру на следующем простом *примере*. Пусть $k = 2$ и $\hat{R}_X = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.

Пусть выполняется равенство $\hat{\boldsymbol{R}}_{\boldsymbol{X}} \boldsymbol{C} = s\boldsymbol{C}$, где $\boldsymbol{C} \in R^2$, $\boldsymbol{C} \neq \boldsymbol{0}$ и s – некоторое число. Перепишем его в виде $\left(\hat{\boldsymbol{R}}_{\boldsymbol{X}} - s\boldsymbol{E} \right) \boldsymbol{C} = \boldsymbol{0}$. Относительно $\boldsymbol{C}$ получаем систему линейных однородных алгебраических уравнений, которая имеет ненулевое решение тогда и только тогда, когда определитель системы $\left| \hat{\boldsymbol{R}}_{\boldsymbol{X}} - s\boldsymbol{E} \right| = 0$. Корни этого уравнения, называемого *характеристическим*, являются собственными числами матрицы $\hat{\boldsymbol{R}}_{\boldsymbol{X}}$. Раскрыв определитель, получим $s^2 - 2s + 0.75 = 0$. Это уравнение второго порядка имеет два корня $s_1 = 1.5$, $s_2 = 0.6$, которые

и будут собственными числами матрицы $\hat{\boldsymbol{R}}_{\boldsymbol{X}}$. Для вычисления соответствующих им ненулевых собственных векторов решаем две вырожденные однородные системы уравнений $\left(\hat{\boldsymbol{R}}_{\boldsymbol{X}} - s_1\boldsymbol{E}\right)\boldsymbol{C} = \boldsymbol{0}$ и $\left(\hat{\boldsymbol{R}}_{\boldsymbol{X}} - s_2\boldsymbol{E}\right)\boldsymbol{C} = \boldsymbol{0}$. Первая из них после перехода от матричной записи к скалярной образует два одинаковых уравнения: $-0.5c_1 + 0.5c_2 = 0,\ 0.5c_1 - 0.5c_2 = 0.$ Оставив одно из них и положив, например, $c_2 = 1,$ найдем первый собственный вектор $\boldsymbol{C}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, соответствующий собственному числу $s_1 = 1.5.$ Поступив аналогичным образом со второй системой уравнений, получим второй собственный вектор, соответствующий числу $s_2 = 0.75,$ например, в виде $\boldsymbol{C}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$

Следует иметь в виду, что каждому собственному числу соответствует неограниченное число собственных векторов, поэтому мы используем слово “например”. Однако нормированный собственный вектор будет с точностью до знака один, и тогда становится несущественным соображение, которым мы руководствуемся при вычислении конкретного собственного вектора. В данном примере нормированными собственными векторами будут векторы $\tilde{\boldsymbol{C}}_1 = \dfrac{\boldsymbol{C}_1}{\left\|\boldsymbol{C}_1\right\|} = \dfrac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix},\ \tilde{\boldsymbol{C}}_2 = \dfrac{\boldsymbol{C}_2}{\left\|\boldsymbol{C}_2\right\|} = \dfrac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix},$ где $\left\|...\right\|$ – евклидова норма вектора. Тогда ортонормированная матрица $\boldsymbol{U}$ оказывается равной $\boldsymbol{U} = \dfrac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. Несложно убедиться в справедливости (2.13):

$$\boldsymbol{U}^T\hat{\boldsymbol{R}}_{\boldsymbol{X}}\boldsymbol{U} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.5 \end{bmatrix}.$$

Из этого элементарного примера видно, что построение ортогональной (ортонормированной) матрицы является трудоемкой операцией, сложность которой резко возрастает при увеличении порядка k. Чтобы облегчить сопутствующие этому процессу “страдания”, нужно использовать возможности современных пакетов прикладных программ, многие из которых имеют специальные встроенные функции для возвращения собственных чисел и векторов матриц. Так, в среде Mathcad для данного примера элементарно получаем

$$R_X := \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \qquad s := \text{eigenvals}(R_X) \qquad U := \text{eigenvecs}(R_X)$$

$$s = \begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix} \qquad U = \begin{pmatrix} 0.707 & 0.707 \\ -0.707 & 0.707 \end{pmatrix} \qquad U^T \cdot R_X \cdot U = \begin{pmatrix} 0.5 & 0 \\ 0 & 1.5 \end{pmatrix}.$$

Обратим внимание на то, что в данном случае компьтер не упрядочил компоненты вектора s в направлении невозрастания и это следует сделать самостоятельно:

$$s := \text{reverse}(s) \qquad s = \begin{pmatrix} 1.5 \\ 0.5 \end{pmatrix} \qquad i := 0..1 \qquad U^{\langle i \rangle} := \text{eigenvec}(R_X, s_i)$$

$$U = \begin{pmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{pmatrix} \qquad U^T \cdot R_X \cdot U = \begin{pmatrix} 1.5 & 0 \\ 0 & 0.5 \end{pmatrix}$$

Полезно напомнить, что если вектор $\boldsymbol{U}^{(i)}$ является собственным вектором, то тому же собственному числу будет соответствовать и вектор $-\boldsymbol{U}^{(i)}$.

Итак, пусть проведена операция (2.13) и получена матрица $\boldsymbol{S}$. В обозначениях равенства (2.9) это означает

$$\boldsymbol{S} = \boldsymbol{U}^T \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{U}. \qquad (2.14)$$

Представим

$$\boldsymbol{S} = \boldsymbol{S}^{0.5}\boldsymbol{S}^{0.5}, \quad \boldsymbol{S}^{0.5} = \begin{bmatrix} \sqrt{s_1} & 0 & \ldots & 0 \\ 0 & \sqrt{s_2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sqrt{s_k} \end{bmatrix}. \qquad (2.15)$$

Тогда равенство (2.14) выполняется, если положить $\boldsymbol{S}^{0.5} = \boldsymbol{U}^T \boldsymbol{A}$, откуда следует

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}^{0.5}. \qquad (2.16)$$

Это и будет искомым выражением для матрицы нагрузок $\boldsymbol{A}$. Таким образом, для построения матрицы нагрузок следует: вычислить собственные числа выборочной корреляционной матрицы $\hat{\boldsymbol{R}}_X$ и упорядочить их в направлении невозрастания $s_1 \geq s_2 \geq \ldots \geq s_k$; для каждого собственного числа s_i, $i = \overline{1,k}$, найти ненулевой нормированный собственный вектор; расположив эти векторы в форме столбцов, постро-

ить ортонормированную матрицу $\boldsymbol{U}$; используя (2.15), (2.16), вычислить матрицу нагрузок $\boldsymbol{A}$.

Полезно обратить внимание на одно характерное свойство матрицы $\boldsymbol{A}$. Если столбцы матриц $\boldsymbol{A}$ и $\boldsymbol{U}$ обозначить символами соответственно $\boldsymbol{A}^{(i)}$, $\boldsymbol{U}^{(i)}$, $i=\overline{1,k}$, то из (2.16) получим

$$\boldsymbol{A}^{(i)} = \sqrt{s_i}\boldsymbol{U}^{(i)},\ i=\overline{1,k}.$$

Так как столбцы матрицы $\boldsymbol{U}$ ортогональны, то и столбцы матрицы $\boldsymbol{A}$ оказываются также *ортогональными*.

2.4. Вклад главных компонентов в суммарную дисперсию

В предыдущем разделе было отмечено, что величина $\sum_{i=1}^{k} a_{i,j}^2$ представляет собой вклад j-го главного компонента $\left(j=\overline{1,k}\right)$ в суммарную дисперсию всех стандартизованных исходных признаков, причем все вклады должны быть ранжированы в смысле выполнения неравенств:

$$\lambda_1 = \sum_{i=1}^{k} a_{i,1}^2 \geq \lambda_2 = \sum_{i=1}^{k} a_{i,2}^2 \geq \lambda_3 = \sum_{i=1}^{k} a_{i,3}^2 \geq \ldots \geq \lambda_k = \sum_{i=1}^{k} a_{i,k}^2 . \qquad (2.17)$$

Сейчас построена матрица $\boldsymbol{A}$ в форме (2.16). Покажем, что использованная при этом построении технология обеспечивает выполнение неравенств (2.17).

Обратившись к соотношению (2.13), запишем

$$Sp\boldsymbol{U}^T\hat{\boldsymbol{R}}_{\boldsymbol{X}}\boldsymbol{U} = Sp\boldsymbol{S} = \sum_{i=1}^{k} s_i .$$

След матрицы $\boldsymbol{U}^T\hat{\boldsymbol{R}}_{\boldsymbol{X}}\boldsymbol{U}$ не изменится, если сомножители $\boldsymbol{U}^T\hat{\boldsymbol{R}}_{\boldsymbol{X}}$ и $\boldsymbol{U}$ поменять местами [18], откуда следует

$$Sp\boldsymbol{U}^T\hat{\boldsymbol{R}}_{\boldsymbol{X}}\boldsymbol{U} = Sp\boldsymbol{U}\boldsymbol{U}^T\hat{\boldsymbol{R}}_{\boldsymbol{X}} = Sp\hat{\boldsymbol{R}}_{\boldsymbol{X}} = k .$$

Сопоставляя два последних выражения, заключаем

$$\sum_{i=1}^{k} s_i = k,$$

т.е. сумма всех собственных чисел матрицы $\hat{\boldsymbol{R}}_{\boldsymbol{X}}$ равняется суммарной дисперсии стандартизованных признаков. С другой стороны, из определения (2.16) матрицы $\boldsymbol{A}$ следует

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{S}^{0.5}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{S}^{0.5} = \boldsymbol{S} = \begin{bmatrix} s_1 & 0 & ... & 0 \\ 0 & s_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & s_k \end{bmatrix}$$

или в развернутом виде

$$\begin{bmatrix} a_{1,1} & a_{2,1} & ... & a_{k,1} \\ a_{1,2} & a_{2,2} & ... & a_{k,2} \\ ... & ... & ... & ... \\ a_{1,k} & a_{2,k} & ... & a_{k,k} \end{bmatrix} \begin{bmatrix} a_{1,1} & a_{1,2} & ... & a_{1,k} \\ a_{2,1} & a_{2,2} & ... & a_{2,k} \\ ... & ... & ... & ... \\ a_{k,1} & a_{k,2} & ... & a_{k,k} \end{bmatrix} = \begin{bmatrix} s_1 & 0 & ... & 0 \\ 0 & s_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & s_k \end{bmatrix},$$

что порождает равенства

$$\sum_{i=1}^{k} a_{i,j}^2 = s_j, \quad j = \overline{1,k}.$$

Но в соответствии с (2.17) величина

$$\sum_{i=1}^{k} a_{i,j}^2 = \lambda_j, \quad j = \overline{1,k},$$

определяет вклад j-го главного компонента в общую дисперсию, и так как оказывается $\lambda_j = s_j, \quad j = \overline{1,k},$ этот вклад определяется j-м собственным числом матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$. Эти числа мы упорядочили в направлении $s_1 \geq s_2 \geq s_3 \geq ... \geq s_k$. Следовательно, наибольший вклад в общую дисперсию вносит первый главный компонент, второй по величине вклад обеспечивает второй главный компонент и так далее, наименьший вклад обусловлен k-м главным компонентом, что изначально закладывалось в идею метода главных компонентов.

Для полного завершения задачи следует найти матрицу $\boldsymbol{F}$ значений главных компонентов на n инспектируемых объектах. Это осуществляется применением выражения (2.5):

$$\boldsymbol{F} = \boldsymbol{A}^{-1}\overline{\boldsymbol{Y}}. \tag{2.18}$$

При решении прикладных задач, основанных на применении главных компонентов, обычно с целью уменьшения размерности задачи используют не все k главных компонентов, а только часть из них в количестве $m < k$, относительный вклад δ_m которых в суммарную

дисиперсию k составляет не ме- нее 0.7 $(\delta_m \geq 0.7)$. Тогда оставляют те главные компоненты, для которых $\frac{\sum_{i=1}^{m} s_i}{k} = \delta_m$.

Главные компоненты, что уже отмечалось, являются гипотетическими и не имеют явного физического, в частности экономического, содержания. Однако при решении прикладных задач естественным является стремление придать им определенный смысл. В ряде случаев это достигается путем проведения следующего дополнительного анализа. Возвратимся к соотношению (2.1) и обсудим вклад для определенности первого главного компонента в систему всех исходных признаков. Очевидно, он определяется элементами первого столбца матрицы нагрузок $\boldsymbol{A}$. Выделим в его составе ряд элементов, индексы которых образуют такое множество G, что

$$\frac{\sum_{i \in G} a_{i,1}^2}{\sum_{i=1}^{k} a_{i,1}^2} = \frac{\sum_{i \in G} a_{i,1}^2}{s_1} \in (0.75, 0.95) .$$

Вошедшим в множество G нагрузкам соответствуют и вполне определенные исходные признаки с конкретным содержанием, которое и определит смысл первого главного компонента. Поступая аналогичным образом со вторым, третьим и т.д. столбцами матрицы $\boldsymbol{A}$, сможем содержательно интерпретировать остальные главные компоненты.

3. ОСНОВЫ ФАКТОРНОГО АНАЛИЗА

3.1.Основное уравнение факторного анализа

Ситуация, исследуемая средствами факторного анализа, в значительной степени близка к той, которая изучается в методе главных компонентов. Некоторое явление характеризуется k признаками, обобщенными в форме вектора $\boldsymbol{X} \in R^k$. Этот вектор стремятся представить как результат линейного преобразования некоторой новой комбинации в общем случае гипотетических величин, но меньшей, чем k, размерностью и обладающей рядом специфичных свойств. Факторный анализ, таким образом, направлен на снижение (редукцию) размерности начального представления проблемы, выявление причин, определяющих свойства исходных признаков, и моделирование их в классе линейных функций, наделенных определенными свойствами.

Итак, пусть, как и в методе главных компонентов, вектор $\boldsymbol{X} = \left[X_1,...,X_k\right]^T$ представляет собой некоторую совокупность признаков, отражающих свойства исследуемого явления и обладающих свойствами, характерными для метода главных компонентов. Но в отличие от модели главных компонентов (2.1) в факторном анализе каждый компонент X_i, $i=\overline{1,k}$, вектора $\boldsymbol{X}$ аппроксимируют линейной комбинацией $m<k$ новых гипотетических признаков, влияющих на все k исходных признаков, и одним индивидуальным признаком, влияющим только на какой-либо конкретный компонент вектора исходных признаков. Условно эта исходная предпосылка факторного анализа отображается записью

$$X_i = \sum_{r=1}^{m} a_{i,r} f_r + d_i \nu_i, \quad i=\overline{1,k}. \tag{3.1}$$

В модели (3.1) величины f_r, $r=\overline{1,m}$, называют *общими* факторами, действующими на все исходные признаки, а параметр $a_{i,r}$ моделирует нагрузку i-го исходного признака на r-й общий фактор. Величина ν_i представляет собой индивидуальный или *характерный* фактор, действующий только на i-й исходный признак с весом d_i $(i=\overline{1,k})$.

Как и в компонентном анализе, вектор $\boldsymbol{X}$ регистрируется на n каких-либо объектах, следствием чего является матрица (2.2) $\overline{\boldsymbol{X}}$, после проведения операций централизации приводящая к матрице централизованных признаков (2.3) $\overline{\boldsymbol{Y}}$. Произвольный ij-й элемент $\overline{y_{i,j}}$ этой матрицы в обозначениях (3.1) может быть представлен выражением

$$\overline{y_{i,j}} = \sum_{r=1}^{m} a_{i,r} f_{r,j} + d_i \nu_{i,j}, \quad i=\overline{1,k}, \quad j=\overline{1,n}, \tag{3.2}$$

где $f_{r,j}$ - значение r-го общего фактора на j-м объекте, $\nu_{i,j}$ - значение i-го характерного фактора на том же объекте. Тогда матрица $\overline{\boldsymbol{Y}}$ приобретает вид

$$\overline{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{F} + \boldsymbol{D}\boldsymbol{\nu}, \tag{3.3}$$

где использованы очевидные обозначения:

$\boldsymbol{A} = \left[a_{i,r}\right] \in R^{k\times m}$, $\boldsymbol{F} = \left[f_{r,j}\right] \in R^{m\times n}$, $\boldsymbol{D} = diag\left[d_i\right] \in R^{k\times k}$, $\boldsymbol{\nu} = \left[\nu_{i,j}\right] \in R^{k\times n}$.

Заметим, что модель (3.3) является *неоднозначной* в том смысле, что если ей удовлетворяют некоторые матрицы $\boldsymbol{A}$, $\boldsymbol{F}$, то этому же

уравнению будут удовлетворять любые матрицы $\tilde{\boldsymbol{A}} = \boldsymbol{AB}$, $\tilde{\boldsymbol{F}} = \boldsymbol{B}^{-1}\boldsymbol{F}$, где $\boldsymbol{B}$ - любая $m \times m$ невырожденная матрица.

В результате проведения факторного анализа необходимо определить рациональное в некотором смысле число m общих факторов, суметь их содержательно интерпретировать, построить матрицы $\boldsymbol{A}, \boldsymbol{F}, \boldsymbol{D}, \boldsymbol{\nu}$. Основополагающей, как и в методе главных компонентов, в факторном анализе является информация, содержащаяся в выборочной корреляционной матрице $\hat{\boldsymbol{R}}_{\boldsymbol{X}}$, которая в случае (3.3) равна

$$\hat{\boldsymbol{R}}_{\boldsymbol{X}} = \frac{1}{n-1}\overline{\boldsymbol{Y}}\,\overline{\boldsymbol{Y}}^{T} =$$

$$= \frac{1}{n-1}\boldsymbol{AFF}^{T}\boldsymbol{A}^{T} + \frac{1}{n-1}\boldsymbol{AF\nu}^{T}\boldsymbol{D} + \frac{1}{n-1}\boldsymbol{D\nu F}^{T}\boldsymbol{A}^{T} + \frac{1}{n-1}\boldsymbol{D\nu\nu}^{T}\boldsymbol{D}.$$

Относительно матриц $\boldsymbol{F}$ и $\boldsymbol{\nu}$ формируют ряд требований, которым эти матрицы должны удовлетворять. Требования к матрице $\boldsymbol{F}$ те же, что и при компонентном методе: центрирование, нормировка, не коррелированность. При выполнении этих условий $\frac{1}{n-1}\boldsymbol{FF}^{T} = \boldsymbol{E} \in R^{m\times m}$. Относительно матрицы $\boldsymbol{\nu}$ выдвигаются аналогичные требования и, как следствие, $\frac{1}{n-1}\boldsymbol{\nu\nu}^{T} = \boldsymbol{E} \in R^{k\times k}$. Дополнительно предполагается ортогональность матриц $\boldsymbol{F}$ и $\boldsymbol{\nu}$ в смысле выполнения ограничений $\boldsymbol{F\nu}^{T} = \boldsymbol{\nu F}^{T} = \boldsymbol{0} \in R^{m\times m}$, где $\boldsymbol{0}$ - нулевая матрица указанной размерности. И тогда следует

$$\hat{\boldsymbol{R}}_{\boldsymbol{X}} = \boldsymbol{AA}^{\boldsymbol{T}} + \boldsymbol{D}^{\boldsymbol{2}}. \tag{3.4}$$

Обозначим $\boldsymbol{AA}^{\boldsymbol{T}} = \hat{\boldsymbol{R}}_{\boldsymbol{h}}$. Используют терминологию: матрицу $\hat{\boldsymbol{R}}_{\boldsymbol{h}}$ называют *редуцированной* выборочной корреляционной матрицей или матрицей *общностей*, матрицу $\boldsymbol{D}^{2}$ - матрицей *характерностей*. Уравнение

$$\boldsymbol{AA}^{\boldsymbol{T}} = \hat{\boldsymbol{R}}_{\boldsymbol{h}} \tag{3.5}$$

часто называется *фундаментальной теоремой факторного анализа.*

Из сопоставления выражений (3.5) и (3.4) следует

$$\hat{\boldsymbol{R}}_h = \hat{\boldsymbol{R}}_{\boldsymbol{X}} - \boldsymbol{D}^2 = \begin{bmatrix} 1-d_1^2 & \hat{r}_{X_1,X_2} & \hat{r}_{X_1,X_3} & \dots & \hat{r}_{X_1,X_k} \\ \hat{r}_{X_2,X_1} & 1-d_2^2 & \hat{r}_{X_2,X_3} & \dots & \hat{r}_{X_2,X_k} \\ \dots & \dots & \dots & \dots & \dots \\ \hat{r}_{X_k,X_1} & \hat{r}_{X_k,X_2} & \hat{r}_{X_k,X_3} & \dots & 1-d_k^2 \end{bmatrix}. \qquad (3.6)$$

Величины $h_i^2 = 1-d_i^2,\ i=\overline{1,k},$ называют *общностями*. Если общности каким-либо образом найдены, то найдены и параметры $d_i,\ i=\overline{1,k},$ т.е. матрица $\boldsymbol{D}$. Матрицу $\boldsymbol{A}$ при этом ищут как решение уравнения (3.5) при некоторых дополнительных ограничениях.

3.2. Построение редуцированной корреляционной матрицы

Из (3.6) следует, что построение матрицы $\hat{\boldsymbol{R}}_{\boldsymbol{h}}$ практически заключается в вычислении общностей $h_i^2,\ i=\overline{1,k}.$ В настоящее время с этой целью используют различные подходы, большинство из которых носит эвристический характер и не подкреплено серьезными теоретическими обоснованиями. Приведем ряд таких подходов.

1. Использование множественного коэффициента корреляции.

Этот метод признается теоретически наиболее обоснованным. В качестве общности h_i^2 принимается квадрат множественного выборочного коэффициента корреляции i-го компонента X_i вектора $\boldsymbol{X}$ с остальными компонентами этого вектора, т.е. в обозначениях (1.55)

$$h_i^2 = \hat{q}_i^2 = 1 - \frac{\left|\hat{\boldsymbol{R}}_{\boldsymbol{X}}\right|}{\hat{R}_{i,i}},\quad i=\overline{1,k}. \qquad (3.7)$$

2. В качестве общности h_i^2 используют наибольший по модулю элемент в i-й строке матрицы $\hat{\boldsymbol{R}}_{\boldsymbol{X}}$, не равный единице,

$$h_i^2 = \max_j \left\{ \left| \hat{r}_{X_i,X_j} \right| \right\},\quad i,j=\overline{1,k},\quad i \neq j. \qquad (3.8)$$

3. В ряде случаев вместо (3.8) используют усредненную величину

$$h_i^2 = \frac{1}{k-1} \sum_{j=1,\, j\neq i}^{k} \left| \hat{r}_{X_i,X_j} \right|,\ i=\overline{1,k}. \qquad (3.9)$$

4. В практике построения общностей h_i^2 используется так называемый метод *триад*, основанный на следующем принципе. В $i-$й строке матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{h}}$ отыскиваются два наибольших по модулю элемента, не равных единице, т.е.

$$\left|\widehat{r}_{X_i,X_q}\right| = \max_{j\neq i}\left\{\left|\widehat{r}_{X_i,X_j}\right|\right\},\ \left|\widehat{r}_{X_i,X_v}\right| = \max_{j\neq i,\, j\neq q}\left\{\left|\widehat{r}_{X_i,X_j}\right|\right\},$$

и выделяется корреляция $\widehat{r}_{X_q,X_v}$ между признаками X_q и X_v. Тогда принимается

$$h_i^2 = \frac{\left|\widehat{r}_{X_i,X_q}\right|\cdot\left|\widehat{r}_{X_i,X_v}\right|}{\left|\widehat{r}_{X_q,X_v}\right|},\quad i=\overline{1,k}. \tag{3.10}$$

3.3. Определение размерности матрицы нагрузок

В основном уравнении факторного анализа (3.5) принимается $\boldsymbol{A}\in R^{k\times m}$, где $k-$количество исходных факторов, $m-$число общих факторов, подлежащее определению. Вообще говоря, размерность m исследователь, проводящий факторный анализ, может задать волевым образом, ориентируясь на интуитивные представления о последствиях своего выбора, или отложить этот выбор на “потом”, получив более емкую для принятия этого решения информацию. Если число m общих факторов устанавливается на начальном этапе факторного анализа, интуицию исследователя можно подкрепить рядом формализованных аргументов. Приведем два характерных из них.

Возвратимся к редуцированной корреляционной матрице $\widehat{\boldsymbol{R}}_h$. В общем случае эта матрица, в отличие от корреляционной матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$, утратила свойства положительной определенности и, следовательно, среди ее собственных чисел λ_i, $i=\overline{1,k}$, есть как вещественные положительные, так и числа других видов. Известно [18] общее свойство

$$\sum_{i=1}^{k}\lambda_i = Sp\widehat{\boldsymbol{R}}_h = \sum_{i=1}^{k}h_i^2\ .$$

Пусть $\lambda_j^*>0,\ j\in J,-$положительные собственные числа и $J-$множество индексов этих чисел. Тогда $\sum_{j\in J}\lambda_j^* > Sp\widehat{\boldsymbol{R}}_h$. Мощность множства J меньше числа k, и на множестве J можем определить такое подмно-

жество J_1 $(J_1 \subset J)$ мощностью $m < k$, что $\sum_{j \in J_1} \lambda_j^* \approx Sp\widehat{\boldsymbol{R}}_h$. Это число m рекомендуется принять за количество общих факторов в модели факторного анализа.

После построения матрицы $\boldsymbol{A}$ (см. ниже) более формализованным при уточнении m оказывается *критерий Бартлетт – Лоулли*. Обсуждаются две гипотезы: $H_0 - m$ достаточная для проведения факторного анализа величина и $H_1 - m$ величина недостаточная. Проверяется нулевая гипотеза. Составляется статистика:

$$\gamma = Q \ln \frac{\left|\boldsymbol{A}\boldsymbol{A}^T\right|}{\left|\widehat{\boldsymbol{R}}_{\boldsymbol{X}}\right|}, \quad Q = n - \frac{1}{6}(2k+5) - \frac{2}{3}m.$$

Статистика γ при справедливости гипотезы H_0 распределена по закону χ^2 с $\nu = \frac{1}{2}\left((k-m)^2 - k - m\right)$ степенями свободы, т.е. $\gamma \square \chi^2(\nu)$.

Тогда, если при доверительной вероятности $1-\alpha$ окажется, $\gamma < u_{1-\alpha}$, где $u_{1-\alpha}$ – квантиль порядка $1-\alpha$ распределения $\chi^2(\nu)$, то с вероятностью $1-\alpha$ найденное число m общих факторов считается достаточным. В противном случае, при $\gamma > u_{1-\alpha}$ с вероятностью ошибиться α принимают решение о недостаточном количестве m и его увеличивают.

3.4. Построение матрицы факторных нагрузок

3.4.1. Метод главных факторов в факторном анализе

При проведении факторного анализа используются различные подходы к построению марицы нагрузок $\boldsymbol{A}$ [19]. Остановимся на методе главных факторов. По своему содержанию он напоминает рассмотренный выше метод главных компонентов, но с принципиальными отличиями: матрица $\boldsymbol{A}$ в факторном анализе не квадратная, а прямоугольная размерностью $k \times m$, $m < k$; поиск этой матрицы осуществляется не на основании корреляционной матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$, а с использованием редуцированной матрицы $\widehat{\boldsymbol{R}}_h$. И хотя уравнения (3.5) и (2.9) очень похожи, первое из этих отличий приводит к тому, что плодотворной идеей диагонализации матрицы $\widehat{\boldsymbol{R}}_h$ с помощью ортогонального преобра-

зования $\boldsymbol{U}$ [см. (2.14)-(2.16)] не удается воспользоваться, так как матрица $\boldsymbol{UA}$ в данном случае не является квадратной и решение вида (2.16) невозможно. Однако некоторую концептуальную направленность анализа сохраняем. В частности, потребуем, чтобы первый фактор вносил наибольший вклад в суммарную дисперсию общностей, второй фактор – следующий по величине вклад и т.д., то есть используем принцип компонентного анализа.

Вычисление первого столбца матрицы нагрузок $\boldsymbol{A}$. В факторном анализе главные факторы, в отличие от метода главных компонентов, не объясняют всю дисперсию стандартизованных исходных признаков, но вклад первого главного фактора определяется первым столбцом матрицы $\boldsymbol{A}$, конкретнее, подобно (2.17), величиной

$$J_1 = \sum_{i=1}^{k} a_{i,1}^2 \ . \tag{3.11}$$

Чтобы вклад первого главного фактора оказался наибольшим, элементы $a_{i,1}$, $i = \overline{1,k}$, первого столбца матрицы $\boldsymbol{A}$ следует выбрать из

условия максимума величины (3.13), но при дополнительном условии, что остальные элементы матрицы $\boldsymbol{A}$ удоволетворяют основному уравнению (3.5). Если равенство (3.5) записать в развернутой форме, получим матричное соотношение, из которого более явно виден смысл ограничений:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ a_{2,1} & a_{2,2} & \dots & a_{2,m} \\ \dots & \dots & \dots & \dots \\ a_{k,1} & a_{k,2} & \dots & a_{k,m} \end{bmatrix} \begin{bmatrix} a_{1,1} & a_{2,1} & \dots & a_{k,1} \\ a_{1,2} & a_{2,2} & \dots & a_{k,2} \\ \dots & \dots & \dots & \dots \\ a_{1,m} & a_{2,m} & \dots & a_{k,m} \end{bmatrix} = \begin{bmatrix} h_1^2 & \widehat{r}_{X_1,X_2} & \dots & \widehat{r}_{X_1,X_k} \\ \widehat{r}_{X_2,X_1} & h_2^2 & \dots & \widehat{r}_{X_2,X_k} \\ \dots & \dots & \dots & \dots \\ \widehat{r}_{X_k,X_1} & \widehat{r}_{X_k,X_2} & \dots & h_k^2 \end{bmatrix}.$$

Таким образом, оптимизационная задача приобретает вид

$$J_1 = \sum_{i=1}^{k} a_{i,1}^2 \to \max_{a_{i,1},\, i=\overline{1,k},} \tag{3.12}$$

при ограничениях

$$\sum_{q=1}^{m} a_{i,q} a_{j,q} = \widehat{r}_{X_i,X_j}, \ i = \overline{1,k}, \ \ j = \overline{1,k}, \tag{3.13}$$

причем $\widehat{r}_{X_i,X_i} = h_i^2$. Решение классической задачи на условный экстремум (3.12), (3.13) ищется традиционным методом неопределенных множителей Лагранжа. С этой целью составляется функция Лагранжа

$$L = \sum_{j=1}^{k} a_{j,1}^2 + \sum_{i=1}^{k}\sum_{\nu=1}^{k} \lambda_{i,\nu}\left(\hat{r}_{X_i,X_\nu} - \sum_{j=1}^{m} a_{i,j}a_{\nu,j}\right), \qquad (3.14)$$

где $\lambda_{i,\nu}$ – неопределенный множитель Лагранжа, и ищутся стационарные точки этой функции. Имеем

$$\frac{\partial L}{\partial a_{s,1}} = 2a_{s,1} + \frac{\partial}{\partial a_{s,1}}\sum_{i=1}^{k}\sum_{\nu=1}^{k}\lambda_{i,\nu}\left(\hat{r}_{X_i,X_\nu} - \sum_{j=1}^{m} a_{i,j}a_{\nu,j}\right) =$$

$$= 2a_{s,1} + \frac{\partial}{\partial a_{s,1}}\sum_{i=1}^{k}\sum_{\nu=1}^{k}\lambda_{i,\nu}\left(-a_{i,1}a_{\nu,1}\right) =$$

$$= 2a_{s,1} + \frac{\partial}{\partial a_{s,1}}\sum_{i=1}^{k}\lambda_{i,s}\left(-a_{i,1}a_{s,1}\right) + \frac{\partial}{\partial a_{s,1}}\sum_{\nu=1}^{k}\lambda_{s,\nu}\left(-a_{s,1}a_{\nu,1}\right) =$$

$$= 2a_{s,1} - 2\sum_{i=1}^{k}\lambda_{i,s}a_{i,1} = 0, \quad s = \overline{1,k}.$$

При этих преобразованиях использовано равенство $\lambda_{i,\nu} = \lambda_{\nu,i}$, так как при перемене местами индексов i и ν ограничения не изменяются. Таким образом, относительно элементов первого столбца матрицы $\boldsymbol{A}$ получаем систему уравнений

$$a_{s,1} - \sum_{i=1}^{k}\lambda_{i,s}a_{i,1} = 0, \ s = \overline{1,k}. \qquad (3.15)$$

Эта система должна решаться совместно с ограничениями (3.13). Путем проведения ряда дополнительных преобразований условия (3.15), (3.13) удается представить в форме линейной системы алгебраических уравнений:

$$\begin{aligned}
&\left(h_1^2 - \lambda\right)a_{1,1} + \hat{r}_{X_1,X_2}a_{2,1} + \hat{r}_{X_1,X_3}a_{3,1} + \ldots + \hat{r}_{X_1,X_k}a_{k,1} = 0\\
&\hat{r}_{X_2,X_1}a_{1,1} + \left(h_2^2 - \lambda\right)a_{2,1} + \hat{r}_{X_2,X_3}a_{3,1} + \ldots + \hat{r}_{X_2,X_k}a_{k,1} = 0\\
&\hat{r}_{X_3,X_1}a_{1,1} + \hat{r}_{X_3,X_2}a_{2,1} + \left(h_3^2 - \lambda\right)a_{3,1} + \ldots + \hat{r}_{X_3,X_k}a_{k,1} = 0 \qquad (3.16)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\hat{r}_{X_K,X_1}a_{1,1} + \hat{r}_{X_k,X_2}a_{2,1} + \hat{r}_{X_k,X_3}a_{3,1} + \ldots + \left(h_k^2 - \lambda\right)a_{k,1} = 0\ ,
\end{aligned}$$

причем

$$\lambda = \sum_{s=1}^{k} a_{s,1}^2, \qquad (3.17)$$

т.е. представляет собой максимизируемую величину (3.12). Обозначим первый столбец матрицы $\boldsymbol{A}$ символом $\boldsymbol{A}^{(1)}$, т.е. $\boldsymbol{A}^{(1)} = \left[a_{1,1} \quad a_{2,1} \quad ... \quad a_{k,1} \right]^T$, и тогда систему (3.16) можем записать в матричном виде:

$$\widehat{\boldsymbol{R}}_h \boldsymbol{A}^{(1)} = \lambda \boldsymbol{A}^{(1)}. \qquad (3.18)$$

Сопоставляя выражение (3.18) с введенными в п. 2.2 понятиями собственного числа и собственного вектора матрицы $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}$, заключаем, что число λ в (3.18) является сбственным числом матрицы $\widehat{\boldsymbol{R}}_h$, а $\boldsymbol{A}^{(1)}$ – соответствующим этому числу собственным вектором. Но в соответствии с (3.17) величина λ является максимизируемой в (3.12) функцией и, следовательно, из всех k корней s_i, $i = \overline{1,k}$, характеристического уравнения $\left| \widehat{\boldsymbol{R}}_h - s\boldsymbol{E} \right| = 0$ величина λ должна быть наибольшим корнем: $\lambda = \underset{i=\overline{1,k}}{max}\left(s_1, \quad s_2, \quad ..., \quad s_k \right)$. Тогда вектор $\boldsymbol{A}^{(1)}$ для обеспечения равенства (3.17), задающего квадрат нормы вектора $\boldsymbol{A}^{(1)}$, находится как

$$\boldsymbol{A}^{(1)} = \frac{\boldsymbol{S}\sqrt{\lambda}}{\left\| \boldsymbol{S} \right\|},$$

где $\boldsymbol{S}$ – любое нетривиальное решение системы $\left(\widehat{\boldsymbol{R}}_h - \lambda \boldsymbol{E} \right) \boldsymbol{S} = \boldsymbol{0}$ и $\left\| \boldsymbol{S} \right\|$ – евклидова норма этого решения. Если вычисления проводятся в системе Mathcad, то результат достигается мгновенно, например:

$$R_h := \begin{pmatrix} 5 & 4 & 1 \\ 4 & 3 & 8 \\ 1 & 8 & 1 \end{pmatrix} \cdot 0.1 \qquad \lambda := \max\left(\text{eigenvals}\left(R_h \right) \right) \qquad \lambda = 1.203$$

$$S := \text{eigenvec}\left(R_h, \lambda \right) \qquad S = \begin{pmatrix} 0.472 \\ 0.693 \\ 0.545 \end{pmatrix} \qquad A^{\langle 1 \rangle} := S \cdot \sqrt{\lambda} \qquad A^{\langle 1 \rangle} = \begin{pmatrix} 0.518 \\ 0.76 \\ 0.598 \end{pmatrix}$$

Вычисление второго столбца $\boldsymbol{A}^{(2)}$ *матрицы нагрузок* $\boldsymbol{A}$. С этой целью снова возвращаемся к основному уравнению (3.5) и из обеих его частей удаляем составляющие, порожденные уже найденными элементами вектора $\boldsymbol{A}^{(1)}$. Из развернутой формы представления этого уравнения обнаруживается, что подлежащие удалению элементы образуют матрицу

$$\begin{bmatrix} a_{1,1}^2 & a_{1,1}a_{2,1} & \dots & a_{1,1}a_{k,1} \\ a_{2,1}a_{1,1} & a_{2,1}^2 & \dots & a_{2,1}a_{k,1} \\ \dots & \dots & \dots & \dots \\ a_{k,1}a_{1,1} & a_{k,1}a_{2,1} & \dots & a_{k,1}^2 \end{bmatrix} = \boldsymbol{A}^{(1)} \left(\boldsymbol{A}^{(1)} \right)^T .$$

Таким образом, при вычислении второго столбца матрицы $\boldsymbol{A}$ объектом анализа оказывается матрица

$$\widehat{\boldsymbol{R}}_h^{(1)} = \widehat{\boldsymbol{R}}_h - \boldsymbol{A}^{(1)} \left(\boldsymbol{A}^{(1)} \right)^T , \qquad (3.19)$$

которую называют матрицей *остатков*. Формально для определения элементов вектора $\boldsymbol{A}^{(2)}$ над матрицей $\widehat{\boldsymbol{R}}_h^{(1)}$ следует осуществить операции, подобные тем, которые выше были проведены над матрицей $\widehat{\boldsymbol{R}}_h$ с целью вычисления вектора $\boldsymbol{A}^{(1)}$. Однако мы попытаемся обойтись без этой “болезненной” процедуры.

Умножив равенство (3.19) справа на $\boldsymbol{A}^{(1)}$, получим

$$\widehat{\boldsymbol{R}}_h^{(1)} \boldsymbol{A}^{(1)} = \widehat{\boldsymbol{R}}_h \boldsymbol{A}^{(1)} - \boldsymbol{A}^{(1)} \left(\boldsymbol{A}^{(1)} \right)^T \boldsymbol{A}^{(1)} = \lambda \boldsymbol{A}^{(1)} - \boldsymbol{A}^{(1)} \lambda = \boldsymbol{0} , \qquad (3.20)$$

где учтено, что $\boldsymbol{A}^{(1)}$ является сбственным вектором матрицы $\widehat{\boldsymbol{R}}_h$, удовлетворяющим условию (3.17). Из (3.20) следует, что вектор $\boldsymbol{A}^{(1)}$ является *собственным вектором и матрицы* $\widehat{\boldsymbol{R}}_h^{(1)}$, *но с равным нулю собственным числом.*

Из теории матриц известно [18], что у всякой вещественной симметрической матрицы собственные векторы, соответствующие различным собственным числам, ортогональны. Именно такими свойствами обладает матрица $\widehat{\boldsymbol{R}}_h$. Пусть $\boldsymbol{A}^* \neq \boldsymbol{A}^{(1)}$ – какой-либо собственный вектор матрицы $\widehat{\boldsymbol{R}}_h$ с обственным числом λ^*. Умножим равенство (3.19) справа на вектор $\boldsymbol{A}^*$ и учтем последствия данной теоремы:

$$\boldsymbol{R}_h^{(1)} \boldsymbol{A}^* = \boldsymbol{R}_h \boldsymbol{A}^* - \boldsymbol{A}^{(1)} \left(\boldsymbol{A}^{(1)} \right)^T \boldsymbol{A}^* = \lambda^* \boldsymbol{A}^* .$$

Из данного равенства следует, что $\boldsymbol{A}^*$ является собственным вектором с собственным числом λ^* и матрцы $\widehat{\boldsymbol{R}}_h$, и матрицы $\widehat{\boldsymbol{R}}_h^{(1)}$. Иными словами, матрицы $\widehat{\boldsymbol{R}}_h$ и $\widehat{\boldsymbol{R}}_h^{(1)}$ имеют общие собственные векторы. Но тогда матрица $\widehat{\boldsymbol{R}}_h^{(1)}$ для вычисления второго столбца $\boldsymbol{A}^{(2)}$ матрицы $\boldsymbol{A}$ становится ненужной и можно ограничиться непосредственно мат-

рицей $\hat{\boldsymbol{R}}_h$. Находим второе по величине собственное значение $\lambda^{(2)}$ матрицы $\hat{\boldsymbol{R}}_h$:

$$\lambda^{(2)} = \max_{s_i \neq \lambda, i=\overline{1,k}} \{s_1,\ s_2, ...,\ s_k\}$$

и соответствующий ему собственный вектор $\boldsymbol{S}^{(2)}$ как любой нетривиальный корень уравнения

$$(\hat{\boldsymbol{R}}_h - \lambda^{(2)}\boldsymbol{E})\boldsymbol{S}^{(2)} = 0. \qquad (3.21)$$

Тогда

$$\boldsymbol{A}^{(2)} = \frac{\boldsymbol{S}^{(2)}\sqrt{\lambda^{(2)}}}{\left\|\boldsymbol{S}^{(2)}\right\|}. \qquad (3.22)$$

Продолжая подобное исследование последующих собственных чисел и соответствующих им собственных векторов матрицы $\hat{\boldsymbol{R}}_h$, находим количество m главных факторов, требующихся для восстановления матрицы корреляций $\hat{\boldsymbol{R}}_h$.

3.5. Построение матрицы общих факторов на объектах

Возвратимся к соотношению (3.3):

$$\overline{\boldsymbol{Y}} = \boldsymbol{AF} + \boldsymbol{D\nu}. \qquad (3.23)$$

В процессе проведенных операций построены матрицы $\boldsymbol{A}$ и $\boldsymbol{D}$. Теперь следует вычислить матрицы $\boldsymbol{F}$ и $\boldsymbol{\nu}$, в которых сосредоточены данные об общих и характерных факторах на n обследуемых объектах. Обозначив символами $\overline{\boldsymbol{Y}}^{(1)}$ и $\boldsymbol{F}^{(1)}$ первые столбцы матриц $\overline{\boldsymbol{Y}}$ и $\boldsymbol{F}$ соответственно, рассмотрим систему уравнений относительно компонентов вектора $\boldsymbol{F}^{(1)}$

$$\overline{\boldsymbol{Y}}^{(1)} = \boldsymbol{AF}^{(1)} + \boldsymbol{\varepsilon}^{(1)}, \qquad (3.24)$$

где под $\boldsymbol{\varepsilon}^{(1)}$ понимается первый столбец матрицы $\boldsymbol{D\nu}$. Система (3.24) состоит из k уравнений и содержит $m+k$ неизвестных компонентов векторов $\boldsymbol{F}^{(1)}$ и $\boldsymbol{\varepsilon}^{(1)}$. Все неизвестные, как определенные решения ситемы (3.24), найти невозможно. Будем компоненты вектора $\boldsymbol{\varepsilon}^{(1)}$ интерпретировать как некие случайные центрированные величины, некоррелированные друг с другом и имеющие одну и ту же дисперсию (так называемые *гомоскедастичные* величины). Сконцентрируем внимание на оценивании компонентов вектора $\boldsymbol{F}^{(1)}$. Так как $k > m$, то этот вектор в общем случае даже при известном векторе $\boldsymbol{\varepsilon}^{(1)}$ из системы (3.24) найти не удается (не выполняется теорема Кронелера – Ко-

пелли). Величину $\overline{\boldsymbol{Y}}^{(1)} - \boldsymbol{A}\boldsymbol{F}^{(1)}$ назовем *невязкой* уравнения (3.24). И под вектором $\boldsymbol{F}^{(1)}$ условимся понимать его значение, минимизирующее квадрат нормы невязки, т.е. соответствующее условию

$$\left\| \overline{\boldsymbol{Y}}^{(1)} - \boldsymbol{A}\boldsymbol{F}^{(1)} \right\|^2 \to \min_{\boldsymbol{F}^{(1)}}. \qquad (3.25)$$

Такое значение вектора $\boldsymbol{F}^{(1)}$ принято называть *псевдорешением* несовместной ситемы уравнений (3.24) [20] или оценкой вектора $\boldsymbol{F}^{(1)}$, найденной по методу наименьших квадратов (*МНК-оценкой*, [21]). В развернутом виде условие (3.25) приобретает форму

$$\left(\overline{\boldsymbol{Y}}^{(1)} - \boldsymbol{A}\boldsymbol{F}^{(1)}\right)^T \left(\overline{\boldsymbol{Y}}^{(1)} - \boldsymbol{A}\boldsymbol{F}^{(1)}\right) \to \min_{\boldsymbol{F}^{(1)}}.$$

Вычислив градиент этой целевой функции по вектору $\boldsymbol{F}^{(1)}$ и приравняв его нулевому вектору, запишем необходимое условие минимума в форме системы уравнений

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{F}^{(1)} = \boldsymbol{A}^T\overline{\boldsymbol{Y}}^{(1)}. \qquad (3.26)$$

Рассмотрим матрицу $\boldsymbol{A}^T\boldsymbol{A} \in R^{m\times m}$. Пусть матрица $\boldsymbol{A}$ является матрицей *полного* ранга, т.е. $rank(\boldsymbol{A}) = \min(m,k) = m$. Тогда, так как [18] $rank(\boldsymbol{A}^T\boldsymbol{A}) = rank(\boldsymbol{A})$, будем иметь $rank(\boldsymbol{A}^T\boldsymbol{A}) = m$ и матрица $\boldsymbol{A}^T\boldsymbol{A}$ в этом случае оказывается невырожденной. Таким образом, если число общих факторов m меньше количества исходных признаков k, то существует обратная матрица $\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}$ и из (3.26) следует искомое псевдорешение

$$\boldsymbol{F}^{(1)} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\overline{\boldsymbol{Y}}^{(1)}. \qquad (3.27)$$

Подобные (3.24) представления можем распространить на последующие пары столбцов $\boldsymbol{F}^{(2)}, \overline{\boldsymbol{Y}}^{(2)}, ..., \boldsymbol{F}^{(n)}, \overline{\boldsymbol{Y}}^{(n)}$ матриц $\boldsymbol{F}$ и $\overline{\boldsymbol{Y}}$, вычисляя по аналогии с (3.27) величины

$$\boldsymbol{F}^{(i)} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\overline{\boldsymbol{Y}}^{(i)},\ i = \overline{2,n}. \qquad (3.28)$$

В результате оказывается построенной матрица общих факторов $\boldsymbol{F}$ на всех n объектах. Затем можем приступить к вычислению матрицы характерных факторов $\boldsymbol{\nu}$. После того как найдена МНК-оценка матрицы $\boldsymbol{F}$, из (3.23) следует

$$\boldsymbol{\nu} = \boldsymbol{D}^{-1}\left(\overline{\boldsymbol{Y}} - \boldsymbol{A}\boldsymbol{F}\right). \qquad (3.29)$$

3.6. Проблема вращения в факторном анализе

Обсуждая соотношение (3.3), мы уже обращали внимание на его неоднозначность, понимаемую в том смысле, что если равенство (3.3) выполняется при некоторых матрицах $\boldsymbol{A}$ и $\boldsymbol{F}$, то оно будет справедливо при любых матрицах $\tilde{\boldsymbol{A}} = \boldsymbol{AT}$, $\tilde{\boldsymbol{F}} = \boldsymbol{T}^{-1}\boldsymbol{F}$, где $\boldsymbol{T}$ - произвольная невырожденная, в частности ортонормированная, матрица надлежащей размерности. В факторном анализе, пользуясь этим обстоятельством, стремятся построить такую матрицу $\tilde{\boldsymbol{A}}$, которая позволит просто и наглядно интерпретировать результаты факторного анализа. Если матрица преобразования $\boldsymbol{T}$ при этом является ортогональной, то соответствующий процесс перехода от $\boldsymbol{A}$ к $\tilde{\boldsymbol{A}}$ называют *ортогональным вращением*. Матрица $\tilde{\boldsymbol{A}}$, получаемая в результате такого вращения, должна обладать рядом свойств, при обретении которых ее по аналогии с используемой в линейной алгебре терминологией называют матрицей *простой структуры*. В частности, такая матрица должна получиться в определенной степени разреженной в том смысле, что каждый ее столбец будет содержать некоторое количество нулевых элементов. Матрицу $\boldsymbol{T}$ обычно выражают как произведение "элементарных" ортогональных матриц, каждая из которых геометрически обеспечивает поворот относительно одной координатной оси. В результате такой декомпозиции, например, поворот на углы α, β, γ против часовой стрелки можно задать в виде

$$\boldsymbol{T} = \boldsymbol{T}_1\boldsymbol{T}_2\boldsymbol{T}_3,$$

$$\boldsymbol{T}_1 = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{T}_2 = \begin{bmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{bmatrix}, \qquad \boldsymbol{T}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & -\sin\gamma \\ 0 & \sin\gamma & \cos\gamma \end{bmatrix}.$$

Матрица простой структуры должна удовлетворять ряду формализованных ограничений. Основные из них приведены, например, в работах [1], [19].

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Выборка 11

Гауссова случайная величина 7
-стандартная 8
Генеральная совокупность 11

Двумерная гауссова плотность 21
Дисперсия случайной величины 10
Дисперсионная модель 54
Доверительная вероятность 15
Доверительный интервал 14

Закон распределения вероятностей
-одномерный 5
-двумерный 18
-k-мерный 34
Значимость оценки 29

Квантиль 6
Коэффициент ковариации 22
--конкордации Кендалла 52
-корреляция 23
-множественный 38

Математическое ожидание случайные величины 9
Матрица ковариционная 25
-корреляционная 25
-неотрицательно определенная 25
-нагрузок 60
-положительно определенная 25
-подстановок 47
Ортогональная 61

-характеристика 69
Метод максимального правдоподобия 12
Многомерные методы 3
Момент начальный 22
-центральный 22

Некоррелированные величины 23
Несмещенная оценка 12

Общие факторы 68
Однородность двух генеральных совокупностей 43
Операция усреднения 19
Оценка несмещенная 12
-состоятельная 12
-эффективная 12
Оценки выборочные 40
Ошибка оценивания 12
Ошибки 1-го рода 29
Ошибки 2-го рода 29

Плотность вероятностей одномерная 7
--двумерная 19
--k-мерная 34
--условная 20

Процентная точка 6

Ранговая корреляция Кендалла 49
-Спирмена 8
Распределение χ^2 8
-Стьюдента 8
-Фишера 8

Регрессия 27
Редуцированная корреляционная матрица 69

Символ усреднения 9
Случайная величина 5
Среднеквадратичное отклонение10

Условная дисперсия 27
Условное математическое ожидание 27

Уровень значимости 15

Характерные факторы 68
Характеристическое уравнение 62
Хотеллинга статистика 42

Центрированная случайная величина 9

Частные коэффициенты корреляции 37

Библиографический список

1. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. - М.: Финансы и статистика, 2003.
2. Симчера В.М. Методы многомерного анализа статистических данных. -М.: Финансы и статистика, 2008.
3. Сошникова Л.А., Тимашевич В.Н. и др. Многомерный статистический анализ в экономике.- М.: ЮНИТИ-ДАН, 1999.
4. Ниворожкина Л.И., Арженовский С.Б. Многомерные статистические методы в экономике. М.: Издательский дом “Дашков и К”, 2009.
5. Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS (+CD-ROM). /Под ред.Орловой И.В. -М.: Вузовский учебник, 2009.
6. Алехин Е.И. Лекции по многомерному статистическому анализу.– ГОУ ВПО ОГУ, 2007.
7. Калинина В.Н., Соловьев В.И. Введение в многомерный статистический анализ.– ГОУ ВПО ГУУ, 2003.
8. Филатов А.Ю. Конспект лекций по многомерным статистическим методам.– ГОУ ВПО ИГУ, 2007.
9. Дронов С.В. Многомерный статистический анализ. ГОУ ВПО АГУ, 2003.
10. Шаланов Н.В. Многомерный статистический анализ. ГОУ ВПО НГУ, 2005.

11. Венцель Е.С. Теория вероятностей. -М.: Наука, 1964.
12. Колмогоров А.Н. Основные понятия теории вероятностей. – М.: Наука, 1974.
13. Рыжик И.М., Градштейн И.С. Таблицы интегралов, сумм, рядов и произведений. – М.: Физматгиз, 1963.
14. Чураков Е.П. Оптимальные и адаптивные системы. – М.: Энергоатомиздат, 1987.
15. Крамер Г. Математические методы статистики. – М.: Мир, 1975.
16. Колемаев В.А., Староверов О.В., Турундаевский В.Б. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1991.
17. Беллман Р. Введение в теорию матриц. – М.: Наука, 1969.
18. Воеводин В.В., Кузнецов Ю.А. Матрицы и вычисления. – М.: Наука, 1984.
19. Иберла К. Факторный анализ. – М.: Статистика, 1980.
20. Певзнер Л.Д., Чураков Е.П. Математические основы теории систем. – М.: Высшая школа, 2009.
21. Чураков Е.П. Математические методы обработки экспериментальных данных в экономике. – М.: Финансы и статистика, 2004.

Оглавление

Введение……3
1. Характеристики и корреляционный анализ многомерных случайных величин……5
1.1. Скалярные случайные величины……5
1.2. Выборочные оценки числовых характеристик скалярных случайных величин……11
1.3. Доверительный интервал для m_X при известной дисперсии D_X ……15
1.4. Доверительный интервал для m_X при неизвестной дисперсии D_X ……17
1.5. Доверительный интервал для дисперсии D_X ……17
1.6. Многомерные (векторные) случайные величины……...18
1.6.1. Двумерные случайные величины……..18
1.6.2. Числовые характеристики двумерных случайных вели-

чин……22

1.6.3. Выборочные оценки числовых характеристик двумерных случайных величин …….....29
1.6.4. k -мерные случайные величины…….35
1.6.5. Выборочные оценки числовых характеристик k -мерных случайных величин…….40
1.7. Примеры корреляционного анализа в системе Mathcad……44
1.7.1 Анализ однородности двух генеральных совокупностей 56
1.8. Ранговая корреляция……58
1.8.1. Матрица подстановок ……58
1.8.2. Коэффициент ранговой корреляции Кендалла……..60
1.8.3. Коэффициент ранговой корреляции Спирмена……62
1.8.4. Коэффициент конкордации Кендалла……64
1.9. Однофакторная дисперсионная модель……66
2. Основы компонетного анализа……...68
2.1. Постановка проблемы……68
2.2. Основное уравнение компонентного анализа……...69
2.3. Вычисление матрицы нагрузок $\boldsymbol{A}$ ……72
2.4. Вклад главных компонентов в суммарную дисперсию……...77
3. Основы факторного анализа……...79
3.1. Основное уравнение факторного анализа……79
3.2. Построение редуцированной корреляционной матрицы…….82
3.3. Определение размерности матрицы нагрузок……...83
3.4. Построение матрицы факторных нагрузок……...84
3.4.1.Метод главных факторов в факторном анализе..……84
3.5. Построение матрицы общих факторов на объектах……89
3.6. Проблема вращения в факторном анализе……91
Предметный указатель……92
Библиографический список……...93

Ч у р а к о в Евгений Павлович

Многомерные статистические методы

Редактор Р.К. Мангутова
Корректор Н.А. Орлова
Подписано в печать 30.09.13. Формат бумаги 60х84 1/16.
Бумага газетная. Печать трафаретная. Усл. печ. л 6,0.
Тираж 50 экз. Заказ
Рязанский государственный радиотехнический университет.
390005, Рязань, ул. Гагарина, 59/1.
Редакционно-издательский центр РГРТУ.