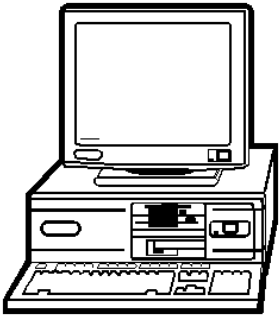
Э.И. Воробьев

МОДЕЛИРОВАНИЕ И АНАЛИЗ СЛОЖНЫХ СИСТЕМ

Учебное пособие



Воронеж 2005

ОГЛАВЛЕНИЕ

1. ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ

Массовые случайные явления обладают свойством статистической устойчивости. Например, частота появления герба при многократном подбрасывании монеты постепенно стабилизируется и приближается к вполне определенному числу — именно к 1/2. Такое же свойство устойчивости обнаруживается и при повторениях любого другого опыта, исход которого представляется заранее неопределенным, случайным. Методы теории вероятностей применимы только к таким экспериментам, которые обладают свойством статистической устойчивости.

Каждый опыт осуществляемого физического эксперимента проводят при одинаковых условиях. Но некоторые факторы при этом не могут быть учтены по объективным причинам. Например, при обработке детали на станке можно поддерживать постоянными скорость и глубину резания, подачу, марку материала и т. д. Однако однородность материала, первоначальные размеры заготовки, вибрации станка и т. д. изменяются в определенных не всегда известных пределах. Поэтому при подобном опыте возможны различные конечные результаты, которые нельзя предсказать до его проведения. Но благодаря постоянным условиям опыта и многократным его повторениям обеспечивается статистическая устойчивость, и в результате можно определить в среднем, какая часть продукции будет годной, а какая нет.

Любой факт (исход, результат), который может появиться или не появиться при проведении опыта, называется в теории вероятностей *случайным событием.* Событие, которое при заданных условиях проведения опытов обязательно произойдет, называется *достоверным.* Если оно не может произойти, его называют *невозможным.*

Количественную меру объективной возможности осуществления некоторого события при фиксированных условиях эксперимента называют вероятностью этого

события. Если при N-кратном повторении опыта рассматриваемое событие A произошло в n_A случаях, то вероятность наступления этого события $P(A)$ определяется выражением

$$P(A) = \lim_{N\to\infty} \frac{\pi A}{N}. \qquad (1)$$

Если событие A достоверное, то $P(A)=1$. Для невозможного события $P(A)=0$. Однако обратные утверждения несостоятельны, так как при $P(A)=1$ не исключено, что в каком-то из опытов событие не произойдет, а равенство $P(A)=0$ вовсе не исключает возможность единичных появлений события A.

События A и B называются статистически независимыми, если вероятность их совместного наступления равна произведению вероятностей этих событий: $P(AB)=P(A)P(B)$. В противном случае события A и B *статистически зависимы.* Вероятность события A, найденная при условии, что осуществилось событие B, называется *условной вероятностью события* A и обозначается $P(A|B)$. Аналогично для события B: $P(B|A)$. Для статистически независимых событий $P(A|B)=P(A)$ и $P(B|A)=P(B)$.

Результаты опытов количественно представляются некоторой *случайной величиной.* Случайной называют физическую величину, принимающую в результате эксперимента то или иное числовое значение, которое в принципе нельзя предсказать исходя из условий проведения опытов. Случайные величины обозначают большими буквами (например X,Y,Z), а их конкретные численные значения (теоретические или наблюдаемые) — соответствующими малыми буквами (x,y,z). Конкретные реализации случайной величины X представляют собой *случайные числа* $x_1, x_2,\ldots,x_N$. Обозначим через R_X множество всех значений, которые может принимать случайная величина X. Множество R_X называют областью значений X. В зависимости от того, как определена область возможных значений, случайные

величины подразделяются на *ограниченные* и *неограниченные, дискретные* и *непрерывные.*

Случайная величина называется ограниченной сверху (снизу) или с обеих сторон, если существует предельное максимальное (минимальное) ее значение или же одновременно и максимальное, и минимальное. На практике большинство случайных величин являются ограниченными. Если подобное ограничение не носит принципиального характера или же границы неизвестны, то можно считать такую случайную величину неограниченной, что иногда удобнее с математической точки зрения.

Случайная величина называется *дискретной,* если ее возможные значения представляют собой дискретный ряд чисел (например, количество выпавших очков при бросании игральной кости; число дефектных изделий в партии и т. п.).

Если область возможных значений случайной величины непрерывна, то ее называют *непрерывной* (например, размер детали, измеренный для партии изготовленных деталей; средняя скорость автомобиля при многократном повторении испытательных заездов на мерном участке пути и т. п.).

Если случайная величина X изменяется в зависимости от одного (y) или нескольких (y_1, y_2, ..., y_m) параметров, то получим *одномерную случайную функцию X(y)* или *случайное поле* $X(y_1, y_2, ..., y_m)$. При $y=t$ случайную функцию называют *случайным процессом X(t).* В теории случайных функций случайный процесс $X(t)$ определяется как случайная функция от неслучайного аргумента — времени t, значения которой при каждом $t \in T$ являются случайными величинами.

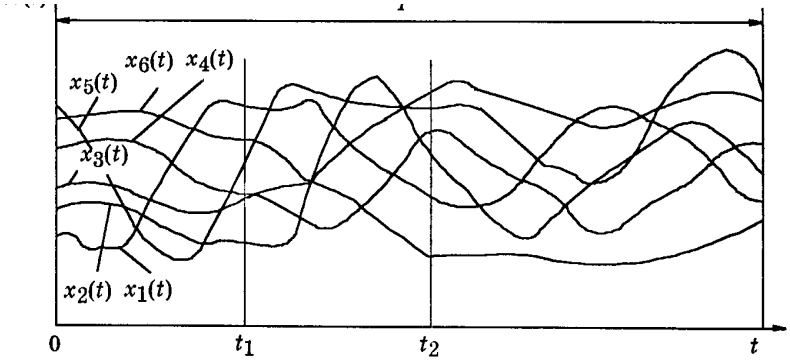


Рис. 1. Ансамбль реализации случайной функции *X(t)*

Конкретный вид, принимаемый случайной функцией в результате опыта, называется *реализацией случайной функции.* Если над технической системой произвести некоторое количество опытов в одинаковых условиях, то получим *ансамбль реализации случайной функции* (рис. 1). Обозначим каждую реализацию случайной функции *X(t)* соответственно номеру опыта $x_1(t)$, $x_2(t),...,x_N(t)$, где N — число опытов. Каждая реализация представляет собой обычную неслучайную функцию. Но так как в каждом опыте функция *X(t)* принимает различный, неизвестный заранее вид, то она является случайной.

Реализации $x_i(t)$, $i = \overline{1,N}$ случайной функции *X(t)* можно получить при одновременном проведении испытаний нескольких одинаковых технических объектов или одного и того же технического объекта в разное время, но в одних и тех же условиях и на одном и том же временном интервале опыта *[0,T]*. Иногда множество N реализации получают расчленением одной реализации достаточно большой длительности путем деления ее на N частей, соответствующих одинаковой длительности процесса T.

Зафиксируем некоторое значение аргумента $t = t_1$ (или $t_1 = t_2$ и т.д.). Очевидно, что в фиксированный момент времени случайная функция *X(t)* превращается в случайную величину

X_1 (или X_2 и т.д.). Полученную случайную величину X_i называют *сечением случайной функции X(t),* соответствующим данному t_i. Если провести сечение ансамбля реализации при t_i (рис. 1), то получим N значений, принятых случайной величиной $X(t)$ в N опытах.

Таким образом, случайная функция совмещает в себе черты случайной величины и функции. Если зафиксировать значение аргумента, она превращается в случайную величину. В результате каждого опыта она превращается в неслучайную функцию.

Воздействия внешней среды на технические системы описываются случайными функциями, а изменения фазовых координат этих систем представляют собой случайные процессы. Поэтому при проектировании таких систем возникает необходимость статистической оценки их характеристик. Кроме того, необходимо моделировать случайные воздействия с заданными характеристиками. При построении экспериментальных факторных моделей также используются методы теории вероятностей и математической статистики. Перейдем к рассмотрению основных характеристик случайных процессов и случайных величин.

1.1. Распределения вероятностей

В теории вероятностей случайные события A рассматривают как точки множества Ω, т. е. $A \in \Omega$, где $\Omega = (\omega_1, \omega_2, ..., \omega_m)$, m — число элементов множества. Случайные события A полностью характеризуются вероятностной мерой $P(A)$, т.е. каждому событию A, как точке некоторого множества Ω, сопоставляется его вероятность $P(A)$.

Для случайных величин и случайных процессов основными вероятностными характеристиками являются *распределения вероятностей*. Они устанавливают связи между реализациями случайной величины X или случайного процесса X(t) и вероятностями их появления.

Наиболее общей формой задания распределения для случайного процесса *X(t)* является *функция распределения F(x,t)* в фиксированный момент времени $t = t_1$. Аргументом этой функции служит реализация *x(t)* случайного процесса в момент времени t . Функция *F(x,t)* определяет вероятность того, что случайный процесс в указанный момент времени t_1 примет значение меньше некоторого уровня *x.* который может варьировать, т.е.

$$F(x,t_1) = P[X(t_1) < x] . \qquad (2)$$

Для случайной величины *X* получим

$$F(x) = P(X \langle x) . \qquad (3)$$

Функции распределения (2) и (3) одномерные. Функция (2) характеризует распределение ординат случайного процесса *X(t)* в отдельные фиксированные моменты времени t_i, а функция (3) — распределение реализации случайной величины *X*.

Основные свойства функции распределения:

1) функция безразмерная и изменяется в пределах $0 \le F(x) \le 1$ для всех *x*;

2) *F(x)* — неубывающая функция *x*: если $x_2 > x_1$, то $F(x_2) > F(x_1)$;

3) для неограниченной случайной величины, т.е. при $-\infty < x < \infty$, $F(-\infty) = P[X \le -\infty] = 0$; $F(+\infty) = P[X \le \infty] = 1$;

4) для ограниченной случайной величины $x_1 \le x \le .x_2$, $F(x_1) = 0$ $F(x_2) = 1$.

На рис. 2, *а* показан график функции *F(x)* для непрерывной случайной величины, а на рис. 2, *б* — для дискретной.

Вероятность того, что случайная величина *X* примет значения в интервале $a \le x \le b$, вычисляют по выражению

$$P(a < X < b) = F(x) = F(b) - F(a) . \qquad (4)$$

Аналогично можно определить вероятность того, что $x > a$:

$$P(X > a) = 1 - F(a) . \qquad (5)$$

Если неограниченно уменьшать интервал, то в пределе, когда $b \to a$, получим

$$P(X = a) = \lim[F(b) - F(a)] = 0 \ . \qquad (6)$$

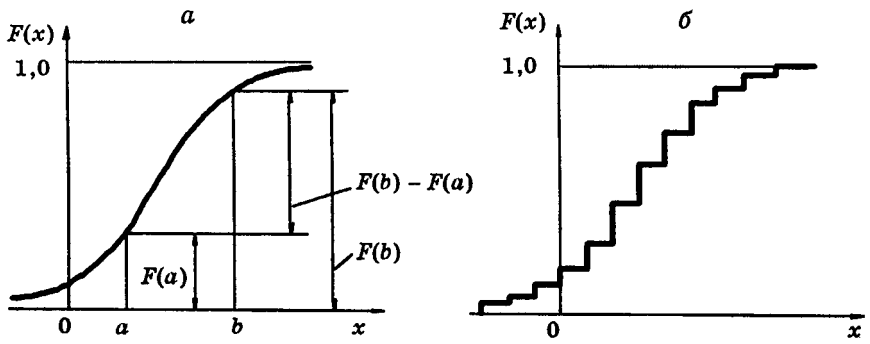


Рис. 2. Графики функции распределения *F(x)* для непрерывной случайной величины (а) и для дискретной случайной величины (б)

Выражение (6) показывает, что вероятность появления события $x = a$ теоретически равна нулю. Однако такое событие при неограниченном числе опытов не следует считать невозможным, но частота его появления будет чрезвычайно малой.

Функции распределения непрерывных случайных величин дифференцируемы по всей области их возможных значений:

$$\frac{dF(x)}{dx} = \lim_{\Delta x \to 0} \frac{P[x \le X \le x + \Delta x]}{\Delta x}; \ \Delta x > 0 . \qquad (7)$$

Обозначим $f(x) = dF(x)/dx$. Функция $f(x)$ приближенно равна отношению вероятности попадания случайной величины внутрь интервала $(x, x + \Delta x)$ к длине интервала Δx. Поэтому

функцию *f(x)* называют *плотностью вероятности.* График функции *f(x)* показан на рис. 3. *Основные ее свойства:*

1) $f(x) \geq 0$;

2) $\int_{-\infty}^{\infty} f(x)dx = 1$;

3) $\int_{-\infty}^{x} f(x)dx = F(x)$.

Зная *f(x),* легко вычислить вероятность нахождения случайной величины внутри любой части области ее возможных значений:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx; \qquad (8)$$

$$P(X \leq a) = \int_{-\infty}^{a} f(x)dx; \qquad (9)$$

$$P(X > b) = \int_{b}^{\infty} f(x)dx. \qquad (10)$$

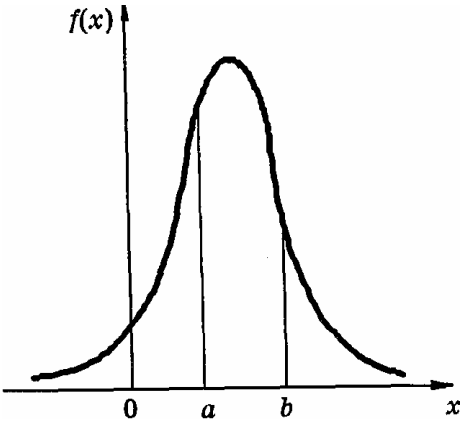


Рис. 3. График плотности распределения *f(x)* непрерывной случайной величины *X*

Согласно этим выражениям определяемые вероятности равны соответствующим площадям под графиком функции *f(x)* (рис. 3).

Функции *F(x,t)* и *f(x,t)* являются простейшими вероятностными характеристиками случайного процесса и определяют одномерное распределение вероятностей в сечении ансамбля реализации при $t = t_i$. Следовательно, они характеризуют случайный процесс изолированно в отдельных его сечениях, не раскрывая взаимной связи между сечениями, т. е. между возможными значениями случайного процесса в различные моменты времени.

Более полными характеристиками случайного процесса будут *функции совместного распределения вероятностей* двух сечений при $t = t_1$ и $t = t_2$ (см. рис.1):

$$F(x_1, x_2; t_1, t_2) = P[X(t_1) \langle x_1; X(t_2) < x_2]; \qquad (11)$$

$$f(x_1, x_2; t_1, t_2) = \partial^2 F(x_1, x_2; t_1, t_2) / \partial x_1 \partial x_2 . \qquad (12)$$

Значения случайного процесса в моменты времени t_1 и t_2 рассматриваются как система двух случайных величин и соотношения (11) и (12) определяют *двумерное распределение вероятности.* При проведении экспериментов на технических объектах часто их результаты представляются не одной, а двумя и более случайными величинами, образующими систему. Распределение системы двух случайных величин X_1 и X_2 описывается функциями:

$$F(x_1, x_2) = P(X_1 < x_1; X_2 < x_2); \qquad (13)$$

$$f(x_1, x_2) = \partial^2 F(x_1, x_2) / \partial x_1 \partial x_2 . \qquad (14)$$

Геометрически функцию $f(x_1, x_2)$ можно изобразить некоторой поверхностью (рис. 4), называемой поверхностью

распределения. Она характеризует *плотность вероятности системы двух случайных величин.* Функцию $f(x_1, x_2)$ называют *двумерной плотностью вероятности.*

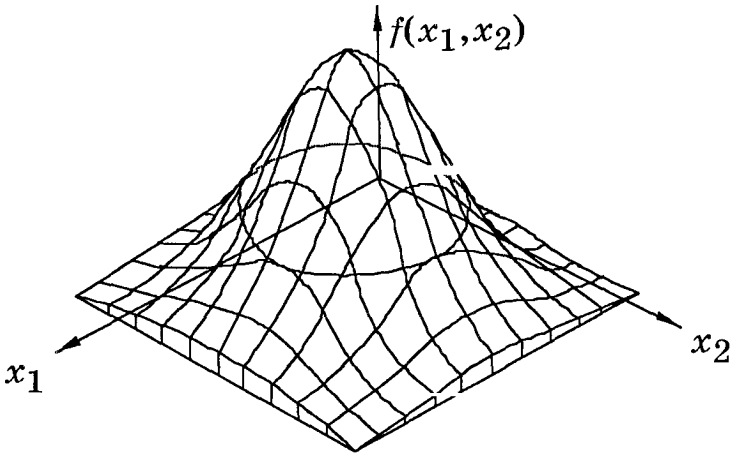


Рис. 4. Графическое отображение двумерной плотности вероятности $f(x_1,x_2)$

Для системы, состоящей из m случайных величин или m сечений ансамбля случайных функций в моменты времени t_1, t_2, ..., t_m получим *m-мерное распределение вероятности.* Многомерное распределение позволяет дать более полное описание случайного процесса, однако получить его сложно, поэтому при исследованиях ограничиваются, как правило, одномерным и двумерным распределениями.

Двумерная плотность вероятности обладает следующими свойствами:

1) $f(x_1, x_2) \geq 0$;

2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1 x_2)dx_1 dx_2 = 1$;

3) $F(x_1, x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(x_1, x_2)dx_1 dx_2$.

По заданным двумерным функциям $F(x_1,x_2)$ или $f(x_1,x_2)$ легко найти функции распределения каждой из случайных величин X_1 и X_2:

$$F(x_1) = F(x_1,+\infty); \quad F(x_2) = F(+\infty, x_2); \qquad (15)$$

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2; \quad f(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_1 \qquad (16)$$

Две случайные величины X_1 и X_2 называются *независимыми,* если их совместная функция распределения является произведением функций распределения этих величин:

$$F(x_1, x_2) = F(x_1)F(x_2); \quad f(x_1, x_2) = f(x_1)f(x_2).$$

Если случайная величина X_2 приняла некоторое конкретное значение a_2, то полученное при этом распределение случайной величины X_1 называют *условным распределением:*

$$F(x_1|a_2) = F(x_1|X_2 = a_2) = \frac{1}{f(a_2)} \int_{-\infty}^{x_1} f(x_1, a_2)dx_1; \qquad (17)$$

$$f(x_1|a_2) = f(x_1|X_2 = a_2) = f(x_1, x_2) / f(a_2). \qquad (18).$$

Аналогично можно определить условные распределения вероятностей X_2 при $X_1 = a_1$.

Условная плотность вероятности $f(x_1|a_2)$ геометрически представляет собой кривую, получающуюся при сечении поверхности $f(.x_1,x_2)$ плоскостью, проходящей через точку a_2 параллельно соответствующей координатной плоскости (см. рис. 4), с последующим умножением каждой из координат на нормирующий множитель $1/f(a_2)$. Необходимость введения нормирующего множителя обусловлена тем, что образующаяся в сечении кривая $f(x_1,a_2)$ не удовлетворяет одному из требований, предъявляемых к функции плотности вероятности, что видно из выражения

$$\int_{-\infty}^{\infty} f(x_1, a_2)dx_1 = f(a_2) \neq 1. \qquad (19)$$

Для независимых случайных величин X_1 и X_2:

$$F(x_1|a_2) = F(x_1); \quad F(x_2|a_1) = F(x_2); \qquad (20)$$

$$f(x_1|a_2) = f(x_1); \quad f(x_2|a_1) = f(x_2). \qquad (21)$$

1.1.1. Числовые вероятностные характеристики

При изучении случайных процессов *X(t)* часто ограничиваются числовыми вероятностными характеристиками — *моментными функциями.* Для случайных величин моментные функции превращаются в обычные числовые характеристики — *моменты распределения вероятностей.*

Моментные функции и моменты бывают начальные и центральные и могут иметь различный порядок. *Начальная моментная функция k-го порядка* определяет математическое ожидание функции $[x(t_i)]^k$ в моменты времени t_i:

$$m_k(x,t_i) = M\left\{\left[X(t_i)\right]^k\right\}, \qquad (22)$$

где *M* — символ математического ожидания.

Математическое ожидание $M[X(t_i)]^k$ есть результат вероятностного усреднения функции $X(t_i)$, т. е. усреднение ее с весом, равным плотности вероятности $f(x,t_i)$. Начальная моментная функция *k*-го порядка вычисляется по формуле

$$m_k(x,t_i) = \int_{-\infty}^{\infty} \left[x(t_i)\right]^k f(x,t_i)dx. \qquad (23)$$

Начальный момент *k*-го порядка случайной величины *X* обозначается $m_k(x) = M[X^k]$. Для непрерывных случайных величин он определяется по формуле

$$m_k(x) = \int_{-\infty}^{\infty} x^k f(x)dx, \qquad (24)$$

а для дискретных

$$m_k(x) = \sum_{i=1}^{n} x_i^k P_i, \qquad (25)$$

где n — число возможных значений случайной величины X; x_i — i-ое значение X; P_i — вероятность, с которой X принимает численное значение x_i.

Начальная моментная функция первого порядка $(k = 1)$ определяет математическое ожидание случайного процесса $X(t)$ в момент времени t_i, т.е. $m_1(x,t_i) = M[X(t_i)]$. Ее значение для сечения ансамбля реализации случайного процесса при t_i вычисляют с использованием выражения

$$m_1(x,t_i) = \int_{-\infty}^{\infty} x(t_i)\, f(x,t_i)dx, \qquad (26)$$

где $f(x, t_i)$ — плотность вероятности случайного процесса в сечении t_i.

Очевидно, что значения $m_1(x, t_i)$ в общем случае различны для разных сечений ансамбля реализации случайного процесса $X(t)$ (штриховая линия на рис. 5). Функцию $m_1(x,t)$ обычно обозначают $m_x(t)$ и называют математическим ожиданием случайного процесса. Можно дать следующую формулировку математического ожидания случайного процесса: *математическим ожиданием случайного процесса* $M[X(t)]$ называется функция времени $m_x(1)$, равная для каждого значения аргумента $t = t_i$ математическому ожиданию случайной величины $X(t_i)$ в сечении t_i ансамбля ее реализации.

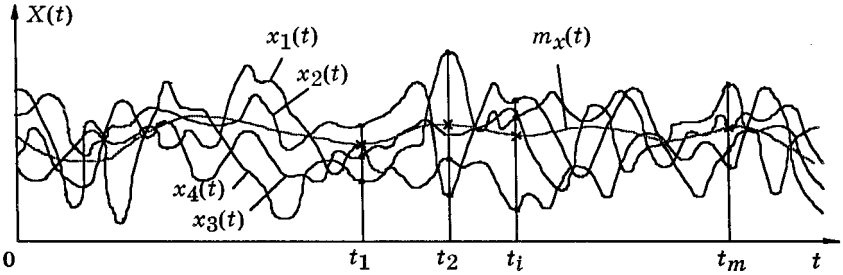


Рис. 5. Ансамбль реализации $x_i(t)$ случайного процесса $X(t)$ и его математическое ожидание $m_x(t)$

Математическим ожиданием случайной величины X является момент первого порядка $m_1(x)$, обозначаемый обычно m_x и вычисляемый по формуле (24) при $k = 1$:

$$m_1(x) = m_x = \int_{-\infty}^{\infty} x f(x) dx . \qquad (27)$$

Начальная моментная функция второго порядка соответствует математическому ожиданию квадрата ординат случайного процесса $X(t)$:

$$m_2(x, t_i) = \int_{-\infty}^{\infty} [x(t_i)]^2 f(x, t_i) dx . \qquad (28)$$

Аналогично для случайной величины X начальный момент второго порядка

$$m_2(x) = \int_{-\infty}^{\infty} x^2 f(x) dx . \qquad (29)$$

У случайных процессов $X(t)$ начальной моментной функцией второго порядка кроме $m_2(x,ti)$ будет и смешанная функция $m_{ll}(x_l,x_2,t_1,t_2)$, определяющая математическое ожидание произведения значений случайного процесса в моменты времени t_1 и t_2:

$$m_{11}(x_1, x_2, t_1, t_2) = M[X(t_1) X(t_2)]. \qquad (30)$$

Функцию (30) называют *ковариационной функцией случайного процесса.* Для системы случайных величин X_1 и X_2 смешанный начальный момент второго порядка называют *ковариационным моментом* и определяют по формуле

$$m_{11}(x_1, x_2) = M[X_1 X_2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f(x_1 x_2) dx_1 dx_2 . \quad (31)$$

Отклонение случайного процесса $X(t)$ от его математического ожидания называют *центрированным случайным процессом* $\overset{o}{X}(t)$:

$$\overset{\circ}{X}(t) = X(t) - m_x(t) \qquad (32)$$

а отклонение случайной величины $\overset{o}{X}(t)$ от ее математического ожидания называют *центрированной случайной величиной* $\overset{o}{X}$*:*

$$\overset{\circ}{X} = X - m_x . \qquad (33)$$

Для центрированных случайных процессов $\overset{o}{X}(t)$ и случайных величин $\overset{o}{X}$ моментные функции $\mu_k(x,t_i)$ и моменты $\mu_k(x)$ *k*-го порядка определяют по выражениям:

$$\mu_k(x,t_i) = \int_{-\infty}^{\infty} [\overset{\circ}{x}(t_i)]^k \, f(x,t_i)dx; \qquad (34)$$

$$\mu_k(x) = \int_{-\infty}^{\infty} \overset{\circ}{x}^k f(x)dx. \qquad (35)$$

Практическое значение при статистическом анализе технических систем имеют центральные моментные функции второго порядка:
для квадрата центрированного процесса

$$\mu_2(\overset{\circ}{x},t_i) = \int_{-\infty}^{\infty} [\overset{\circ}{x}(t_i)]^2 \, f(x,t_i)dx; \qquad (36)$$

смешанная центральная моментная функция (математическое ожидание произведения центрированных значений случайного процесса)

$$\mu_{11}(\overset{\circ}{x}_1, \overset{\circ}{x}_2, t_1, t_2) = M[\overset{\circ}{X}(t_1)\overset{\circ}{X}(t_2)] = \int_{-\infty}^{\infty} \overset{\circ}{x}(t_1)\overset{\circ}{x}(t_2) f(x_1, x_2, t_1, t_2)dx_1 dx_2. \qquad (37)$$

Функцию $\mu_2(x,t_i)$ называют *дисперсией случайного процесса X(t).* Дисперсия характеризует разброс возможных реализаций случайного процесса относительно функции математического ожидания $m_x(t)$.

Дисперсию случайного процесса обозначают $D_x(t)$ или $\sigma_x^2(t)$, где $\sigma_x(t)$ — *среднее квадратическое отклонение случайного процесса:*

$$\sigma_x(t) = \sqrt{D_x(t).} \qquad (38)$$

Дисперсия $D_x(t)$ случайного процесса представляет собой функцию времени, значение которой в момент времени $t = t_i$ равно дисперсии случайной величины $X(t_i)$ в сечении t_i ансамбля реализации.

Смешанную центральную моментную функцию $\mu_{11}(x_1, x_2, t_1, t_2)$ называют *корреляционной функцией случайного процесса* и обозначают $R_x(t_1,t_2)$. Она характеризует степень линейной связи *(корреляцию)* между значениями случайного процесса в различные моменты времени. Следует иметь в виду, что в выражении (37) для $R_x(t_1,t_2)$ оба момента времени t_1 и t_2 рассматриваются в любом сочетании всех возможных текущих значений аргумента t случайного процесса.

Во многих случаях удобнее пользоваться нормированной корреляционной функцией

$$\rho_x(t_1,t_2) = R_x(t_1,t_2)/[\sigma_x(t_1)\sigma_x(t_2)]. \qquad (39)$$

Соотношение (39) представляет собой *коэффициент корреляции* между случайными величинами в сечениях t_1 и t_2. При $t_1 = t_2 = t$ получаем $\rho_x(t,t)=1$, так как значение корреляционной функции в этом случае равно дисперсии: $R_x(t,t) = D_x(t) = \sigma_x^2(t)$.

Для случайной величины X центральный момент первого порядка $\mu_1(\overset{\circ}{x}) = 0$, а центральный момент второго порядка $\mu_2 = (\overset{\circ}{x}) = 0$ представляет собой дисперсию D_x, характеризующую разброс (рассеивание) реализации x_i случайной величины X относительно ее математического ожидания m_x. Дисперсия случайной величины X вычисляется по формуле

$$D_x = \int_{-\infty}^{\infty} \overset{\circ}{x}^2 f(x)dx. \qquad (40)$$

Среднее квадратическое отклонение случайной величины $\sigma_x = \sqrt{D_x}$ называют также *стандартом случайной величины.* В качестве относительной меры рассеивания используют *коэффициент вариации* V_x, измеряемый в процентах:

$$V_x = 100\sigma_x / m_x, \%. \qquad (41)$$

Смешанный центральный момент второго порядка системы двух случайных величин X_1 и X_2 называют *корреляционным моментом* этих величин и обозначают $K_{x_1x_2}$.

$$\mu_{11}(\overset{\circ}{x}_1, \overset{\circ}{x}_2, t_1, t_2) = K_{x_1x_2} = M[\overset{\circ}{X}_1 \overset{\circ}{X}_2] = \\ = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overset{\circ}{x}_1 \overset{\circ}{x}_2 f(x_1, x_2)dx_1dx_2. \qquad (42)$$

В качестве меры взаимозависимости случайных величин X_1 и X_2 используют безразмерную величину, называемую *коэффициентом корреляции*

$$\rho_{x_1x_2} = \frac{K_{x_1x_2}}{\sigma_{x_1}\sigma_{x_2}}. \qquad (43)$$

Коэффициент корреляции может принимать значения в диапазоне $-1 \le \rho_{x_1 x_2} \le 1$. Он определяет степень линейной связи между X_1 и X_2. Знак $\rho_{x_1 x_2}$ зависит от вида линейной связи. При положительном $\rho_{x_1 x_2}$ увеличение одной из случайных величин приводит к возрастанию другой, а при отрицательном, наоборот, к уменьшению. При $\rho_{x_1 x_2} = 0$ случайные величины не коррелированы. Для независимых случайных величин $\rho_{x_1 x_2} = 0$, и, следовательно, они относятся к некоррелированным величинам. Обратное утверждение в общем случае неверно: X_1 и X_2 могут быть связаны даже не статистически, а чисто функционально и все же иметь нулевой коэффициент корреляции. При этом, конечно, указанная функциональная связь должна быть нелинейной. Если $\rho_{x_1 x_2} = 1$, то X_1 и X_2 линейно связаны, т.е. $X_2 = aX_1 + b$.

Для оценки вида графиков характеристик распределения вероятностей случайной величины используют третий и четвертый центральные моменты, вычисляемые по формулам:

$$\mu_3(x) = \int_{-\infty}^{\infty} \overset{\circ}{x}^3 f(x)dx;$$
$$\mu_4(x) = \int_{-\infty}^{\infty} \overset{\circ}{x}^4 f(x)dx. \qquad (44)$$

Третий центральный момент $\mu_3(x)$ определяет асимметрию графика характеристики распределения, показателем которой служит *коэффициент асимметрии*

$$A_x = \mu_3(x)/\sigma_x^3, \qquad (45)$$

а четвертый центральный момент $\mu_4(x)$ характеризует степень заостренности (крутости) графика характеристики

распределения, показатель которой называют *коэффициентом эксцесса*

$$E_x = \mu_4(x)/\sigma_x^4 - 3. \qquad (46)$$

Статистической характеристикой связи двух случайных процессов *X(t)* и *Y(t)* является *взаимная корреляционная функция* $R_{xy}(t_1,t_2)$. Она представляет собой математическое ожидание произведения центрированных случайных процессов $\overset{o}{X}(t)$ и $\overset{o}{Y}(t)$ в моменты времени t_1 и t_2:

$$R_{xy}(t_1,t_2) = M[\overset{\circ}{X}(t)\overset{\circ}{Y}(t)]. \qquad (47)$$

Эта функция характеризует степень связи между сечением процесса *X(t)* при $t = t_1$ и сечением процесса *Y(t)* при $t = t_2$. В отличие от корреляционной функции $R_x(t_1,t_2)$ она несет в себе некоторую информацию о среднем фазовом сдвиге случайных процессов $\overset{o}{X}(t)$ и $\overset{o}{Y}(t)$.

Нормированная взаимная корреляционная функция находится из соотношения

$$\rho_{xy}(t_1,t_2) = R_{xy}(t_1,t_2)/[\sigma_x(t_1)\sigma_y(t_2)]. \qquad (49)$$

В заключение отметим, что все вероятностные характеристики являются неслучайными числовыми характеристиками случайных процессов и случайных величин.

1.1.2. Теоретические распределения вероятностей

При статистическом анализе технических систем используют различные теоретические распределения вероятностей случайных процессов (законы распределения). Для непрерывных случайных процессов наиболее часто употребляют нормальное распределение, распределение

Пирсона, гамма-распределение, экспоненциальное распределение. Для дискретных случайных величин используют биномиальное распределение и распределение Пуассона.

Нормальное распределение (закон Гаусса) находит самое широкое практическое применение. Главная особенность, выделяющая нормальное распределение среди других законов, состоит в том, что оно является предельной формой многих распределений.

Одномерная плотность вероятности нормального распределения случайной величины X определяется выражением

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left[-\frac{(x - m_x)^2}{2\sigma_x^2} \right]. \qquad (50)$$

Из выражения (50) следует, что нормальное распределение полностью определяется двумя числовыми характеристиками: математическим ожиданием m_x и дисперсией $\mathrm{D_x} = \sigma_x^2$. Функция распределения

$$F(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{(x - m_x)^2}{2\sigma_x^2} \right] dx, \qquad (51)$$

или

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{x} \exp(-0{,}5u^2) du, \qquad (52)$$

где u — нормированное значение центрированной случайной величины, выраженное в долях σ_x:

$$u = (x - m_x)/\sigma_x. \qquad (53)$$

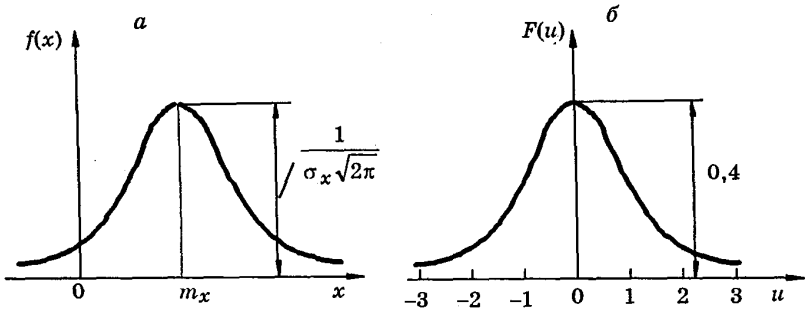


Рис. 6. Графики плотности вероятности нормального распределения случайной величины X (а) и нормированного нормального распределения (б)

Функцию $$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-0,5u^2} \qquad (54)$$

называют *плотностью вероятности нормированного нормального распределения.* На рис. 6, *а* показан график функции f*(x)*, а на рис. 6, *б* — график функции $\varphi(u)$. Влияние математического ожидания m_x и среднего квадратического отклонения σ_x; нормально распределенной случайной величины *X* на вид графика *f(x)* отображено на рис.7. Изменение σ_x при m_x = *const* приводит к перемещению графика *f(x)* вдоль оси *x,* не изменяя его формы (кривые 1,2,3). Изменение σ_x при m_x = const (кривые 2,4,5) приводит к изменению масштабов графика по обеим координатным осям, и форма его изменяется. Однако во всех случаях

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Интегралы, входящие в формулы (51) и (52), не выражаются через элементарные функции. Поэтому для вычисления *F(x)* пользуются таблицами функции

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_0^u \exp(-0{,}5u^2)du, \qquad (55)$$

которая называется *функцией Лапласа (интегралом вероятностей).* Значения функции Лапласа *Ф(u)* приведены в табл. 1 приложения.

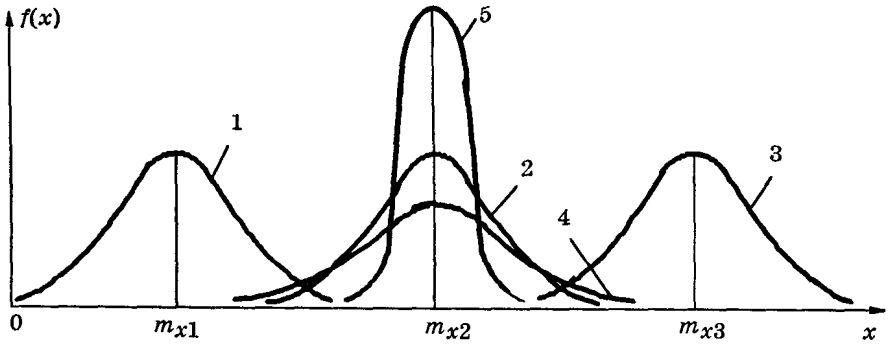


Рис. 7. Влияние параметров m_x и σ_x на график плотности вероятности нормального распределения ($\sigma_{x_5} < \sigma_{x_2} < \sigma_{x_4}$)

Определим вероятность попадания случайной величины *X* в интервал $a < X < b$.

В соответствии с (8) находим

$$P(a < X < b) = \int_a^b f(x)dx = \frac{1}{\sigma_x\sqrt{2\pi}} \int_a^b \exp\left[-\frac{(x-m_x)^2}{2\sigma_x^2}\right]dx.$$

Если ввести нормированную переменную *u* согласно выражению (53), то

$$P(a < X < b) = \Phi(u_2) - \Phi(u_1), \qquad (56)$$

где $u_1 = (a — m_x)/\sigma_x$; $u_2 = (b — m_x)/\sigma_x$.

Отметим, что функция $\Phi(u)$ нечетная, т.е. $\Phi(-u) = -\Phi(u)$, и, кроме того, $\Phi(0) = 0$; $\Phi(-\infty) = -0{,}5$; $\Phi(\infty) = 0{,}5$.

Выражение (56) позволяет найти вероятность отклонения нормально распределенной случайной величины X от ее среднего значения m_x на заданную величину. Найдем вероятность отклонения X от m_x на величину $\pm 3\sigma_x$. Согласно выражению (56):

$$P[m_x - 3\sigma_x) < X < (m_x + 3\sigma_x)] = \Phi(u_2) - \Phi(u_1).$$

Так как в данном случае, согласно (53), $u_1 = -3$; $u_2 = 3$, то $\Phi(u_2) - \Phi(u_2) = 2\Phi(3)$. По табл. 1 приложения находим $2\Phi(3) = 2 \cdot 0{,}4987 \approx 0{,}997$. Полученный результат отражает известное правило «трех сигм», которое гласит, что для нормально распределенной случайной величины отклонения ее от математического ожидания практически не превосходят $3\sigma_x$.

В силу симметричности графика плотности вероятностей нормального распределения (рис. 6, а и б) все нечетные центральные моменты равны нулю, а $\mu_4(x) = 3\sigma_x^4$. В результате коэффициент асимметрии A_x и коэффициент эксцесса E_x в соответствии с выражениями (45) и (45) равны нулю.

Двумерная плотность вероятности $f(x, y)$ нормального распределения системы двух случайных величин X и Y в общем случае имеет вид

$$f(x,y) = (2\pi 2_x \sigma_y \sqrt{1-\rho_{xy}})^{-1} \times \qquad (57)$$

$$\times \exp\left\{ -\frac{1}{2(1-\rho_{xy}^2)} \left[\frac{(x-m_x)^2}{\sigma_x^2} - \frac{2\rho_{xy}(x-m_x)(y-m_y)}{\sigma_x \sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2} \right] \right\}.$$

Для некоррелированных величин $\rho_{xy} = 0$ и двумерная плотность нормального распределения будет равна

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{(x-m_x)^2}{2\sigma_x^2}-\frac{(y-m_y)^2}{2\sigma_y^2}\right], \quad (58)$$

или

$$f(x,y) = f(x)f(y), \quad (59)$$

где $f(x)$ и $f(y)$ — одномерные плотности вероятностей случайных величин X и Y.

Логарифмически нормальное распределение описывает такие случайные величины, для которых нормально распределена не сама величина X, а ее логарифмы (десятичный или натуральный), т.е. $Y = lgX$ (или $Y = lnX$), причем $0 < X < \infty$. Плотность вероятности десятичного логарифма случайной величины определяется выражением

$$f(x)=\begin{cases}0 & при\ x\le 0;\\ (M\sigma_y x\sqrt{2\pi})^{-1}e^{-0{,}5z^2} & при\ x>0,\end{cases} \quad (60)$$

где $z = (lgx - lgx_0)/\sigma_y$; $M = (lge)^{-1} = 2{,}303$; e — основание натуральных логарифмов.

Среднее значение случайной величины и ее дисперсия вычисляются по формулам:

$$m_x = x_0 e^{2{,}65\sigma_y^2}; \quad (61)$$

$$\sigma_x^2 = m_x^2(m_x^2/x_0^2 - 1), \quad (62)$$

причем $m_y = lgx_0$.

Распределение (60) используется в теории надежности для описания времени безотказной работы технических объектов. При малых σ_y ($\sigma_y < 0{,}1...0{,}13$) распределение (60) близко к нормальному.

При $Y = lnX$ в приведенных формулах $M = 1$ и десятичные логарифмы заменяются натуральными.

Распределение Пирсона находит широкое применение в математической статистике и теории надежности. Его

используют для оценки согласованности экспериментальных распределений с теоретическими. Распределением Пирсона с k степенями свободы называют распределение суммы квадратов

$\chi^2 = x_1^2 + x_2^2 + \ldots + x_k^2$ независимых случайных величин, каждая из которых имеет нормальное распределение с $m_x = 0$ и $\sigma_x^2 = 1$. Плотность вероятности распределения Пирсона

$$f(x) = \begin{cases} 0 & при\ x \le 0; \\ \dfrac{e^{-0,5x} x^{0,5k-1}}{2^{0,5k}\,\Gamma(0,5k)} & при\ x > 0, \end{cases} \qquad (63)$$

где $\Gamma(0,5k)$ — гамма-функция, значения которой приводятся к таблицам; k —значение случайной величины χ^2.

Моменты распределения Пирсона: $m_x = k$; $\sigma_x^2 = 2k$ а коэффициенты асимметрии и эксцесса $A_x = \sqrt{8/k}$; $E_x = 12/k$.

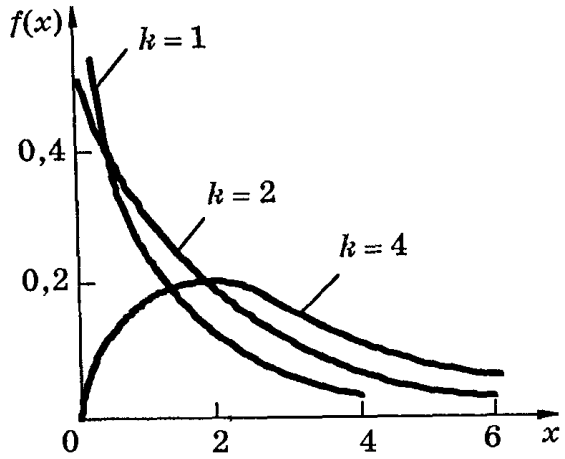


Рис. 8. Графики плотности вероятности распределения Пирсона

С увеличением k распределение χ^2 приближается к нормальному (рис. 8).

Экспоненциальное распределение используется в теории надежности и теории массового обслуживания. Оно определяется одним параметром λ, который называют *интенсивностью потока событий.* Плотность вероятности и функция распределения определяются выражениями:

$$f(x)=\begin{vmatrix} 0 & при\ x\le 0; \\ \lambda e^{-\lambda x} & при\ x>0, \end{vmatrix} \qquad (64)$$

$$F(x)=1-e^{-\lambda x}. \qquad (65)$$

При простейшем потоке отказов значение $e^{-\lambda x}$ определяет вероятность безотказной работы в промежутке $(0,x)$. Моменты экспоненциального распределения $m_x = \sigma_x = \lambda^{-1}$, а коэффициенты асимметрии и эксцесса $A_x = 2$; $E_x = 6$. График функции $f(x)$ показан на рис. 9.

Гамма-распределение представляет собой распределение суммы независимых случайных величин, каждая из которых распределена по экспоненциальному закону. Плотность вероятности (рис. 10)

$$f(x)=\begin{vmatrix} 0 & при\ x\le 0; \\ \dfrac{a^b x^{b-1}}{\Gamma(b)} e^{-ax} & при\ x>0,\ a>0,\ b>0. \end{vmatrix} \qquad (66)$$

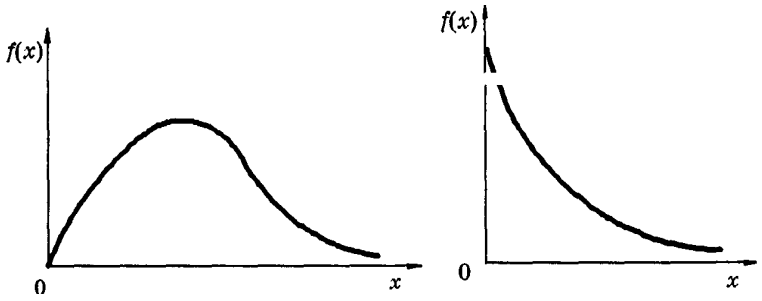


Рис. 9. График плотности вероятности гамма-распределения

Рис 10. График плотности вероятности экспотенциального распределения

Постоянные a и b в выражении (66) являются параметрами гамма-распределения и определяют все его числовые характеристики: $m_x = b/a$; $\sigma_x = \sqrt{b}$; $A_x = 2/\sqrt{b}$; $E_x = 6/b$. Если в выражении (66) положить $2b = k$ и $a = 2$, то получим распределение (63), а при $b = 1$ и $a = \lambda$ — распределение (64).

Биномиальное распределение используется для описания дискретных случайных величин. Вероятность того, что событие A осуществится ровно x раз при N испытаниях, определяется по формуле биномиального распределения:

$$F(x,N) = C_N^x P^x q^{N-x} \quad (x = \overline{0,N}), \qquad (67)$$

где $q = 1-P$ — вероятность неосуществления события A в каждом опыте; C_N^x — число сочетаний из N элементов по x элементов:

$$C_N^x = \frac{N!}{x!(N-x)!}.$$

Математическое ожидание $m_x = NP$, дисперсия $\sigma_x^2 = NPq$.

Распределение Пуассона является предельным для биномиального распределения. При неограниченном увеличении N и уменьшении P так, что при этом $NP = \lambda = const$, получим

$$F(x,\lambda) = \lambda^x e^{-\lambda} / x!. \qquad (68)$$

Величина λ является параметром распределения Пуассона. Выражение (68) описывает распределение числа x случайных событий в каком-либо интервале времени, если можно считать, что вероятность наступления события за интервал Δt пропорциональна этому интервалу и события в разные

моменты времени независимы. Математическое ожидание m_x и дисперсия σ_x^2 распределения равны λ. Значения функции $F(x,\lambda)$ приводятся в таблицах.

Равномерное распределение используют главным образом при моделировании случайных величин. На основе этого распределения составлены таблицы случайных чисел, используемые для решения разных практических задач: случайного взятия проб из каких-либо партий, проведения опытов в случайной последовательности и т. д. (см. табл. 7 приложения).

Вероятность попадания равномерно распределенной случайной величины X в любые равные между собой интервалы, принадлежащие области возможных численных значении X, одна и та же. Это означает, что плотность вероятности случайной величины X постоянна и задается выражением

$$f(x) = \begin{cases} \dfrac{1}{b-a} & при\ a \le x \le b; \\ 0 & при\ x < a,\ x > b, \end{cases} \qquad (69)$$

где a и b — произвольные вещественные числа, являющиеся границами области возможных значений X. Функция распределения

$$f(x) = \begin{cases} \dfrac{x-a}{b-a} & при\ a \le x \le b; \\ 0 & при\ x < a; \\ 1 & при\ x > b. \end{cases} \qquad (70)$$

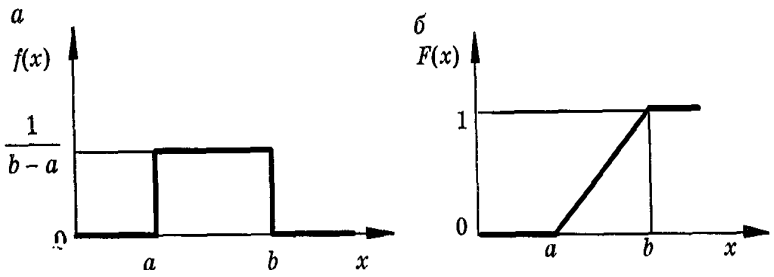


Рис. 11. Графики плотности вероятности *f(x) (а)* и функции распределения *F(x) (б)* для равномерно распределенной случайной величины в интервале *a < X < b*

Графики функций *f(x)* и *F(x)* приведены на рис. 11. Математическое ожидание и дисперсия равномерно распределенной случайной величины *X* определяются по формулам:

$$m_x = (a+b)/2; \qquad (71)$$

$$\sigma_x^2 = (b-a)^2/12. \qquad (72)$$

Обычно в теории и практических приложениях используют случайные числа, равномерно распределенные на интервале *[0, 1]*. В этом случае $m_x = 1/2$, $\sigma_x^2 = 1/12$.

1.2. Моделирование случайных величин

При анализе процессов функционирования вероятностных технических систем возникает необходимость моделирования случайных величин и случайных процессов с заданными вероятностными характеристиками. Так как анализ функционирования технической системы на ЭВМ осуществляется численными методами на основе дискретных

математических моделей, то внешние воздействия на систему необходимо представить в виде некоторой непрерывной последовательности случайных чисел. Рассмотрим способы формирования такой последовательности случайных чисел с заданными вероятностными характеристиками. На практике используют три основных способа генерации случайных чисел: аппаратный (физический), табличный (файловый) и алгоритмический (программный). Наибольшее применение при моделировании технических систем находит алгоритмический способ.

Числа, полученные алгоритмическим способом, реализующим определенные формулы, являются *псевдослучайными.* Однако это не означает, что они некачественные, т. е. неслучайные или случайные зависимые. Качество случайных чисел проверяют соответствующими критериями случайности, независимости и др.

Рассмотрим коротко способ получения псевдослучайных чисел с различными распределениями вероятностей. Как уже отмечалось, исходной базой при этом служат равномерно распределенные в интервале *[0,1]* случайные числа x^R. Для получения таких чисел существует несколько алгоритмов. Один из них описывается в виде рекуррентного соотношения

$$x_{i+1}^R = \lambda x_i^R + \mu (\mathrm{mod}\, M), \tag{73}$$

где λ, μ, M — неотрицательные целые числа.

Случайную величину *x(a,b)*, равномерно распределенную на интервале *[a,b]*, получают на основе выражения

$$x(a,b) = (b-a)x^R + a. \tag{74}$$

Для преобразования равномерно распределенных случайных чисел в случайные числа с заданным распределением вероятностей существует общее правило. Значение функции распределения *F(x)* для равномерно распределенных случайных чисел в интервале *[0, 1]* можно рассматривать как значение случайной величины x^R, т.е.

$$F(x) = x^R. \qquad (75)$$

Решение этого уравнения

$$x = F^{-1}(x^R),$$

где $F^{-1}(x^R)$ — функция, обратная функции $F(x)$, являющаяся случайным числом из совокупности случайных чисел с плотностью вероятности $f(x)$.

Пусть необходимо сформировать случайные числа x^E с экспоненциальным распределением, параметр которого λ. Плотность вероятности этого распределения соответствует выражению (64), а функция распределения — выражению (65). Подставим F(x) из (65) в уравнение (75):

$$1 - e^{-\lambda x^E} = x^R.$$

Решение этого уравнения имеет вид

$$x^E = -\ln[(1 - x^R)]/\lambda. \qquad (76)$$

Так как величина $(1 - x^R)$ также равномерно распределена на интервале $[0,1]$, то ее можно заменить на x^R и использовать выражение

$$x^E = -\ln x^R / \lambda. \qquad (77)$$

Аналогично получают формулу для генерирования случайных чисел x^Γ, соответствующих гамма-распределению

$$x^\Gamma = -\frac{1}{a}\sum_{i=1}^{b}\ln(1 - x_i^R). \qquad (78)$$

При моделировании нормально распределенной случайной величины используют центральную предельную теорему, согласно которой распределение суммы n одинаково распределенных независимых случайных величин $X_1, X_2, \ldots, X_n$ при неограниченном возрастании n неограниченно стремится к нормальному распределению. При $n \geq 8$ распределение этой суммы может считаться нормальным с вероятностью $P > 0{,}95$. Используя реализации случайной равномерно распределенной величины X^R, можно составить выражение

для определения случайной величины X^{N_0}, имеющей нормальное распределение с параметрами $m_x = 0$ и $\sigma_x = 1$:

$$x^{N_0} = \sum_{i=1}^{12} x_i^R - 6. \qquad (79)$$

На основе формулы (79) из двенадцати случайных равномерно распределенных чисел x_i^R получается одно случайное число x^{N_0} новой совокупности, представляющей собой случайные числа с нормальным распределением.

Часто для получения случайных чисел x^{N_0} применяют формулу

$$x^{N_0} = 0{,}774596\left(\sum_{i=1}^{20} x_i^R - 10\right). \qquad (80)$$

Случайные нормально распределенные числа с заданными параметрами m_x и σ_x получают из числа x^{N_0} по формуле

$$x^N = m_x + \sigma_x x^{N_0}. \qquad (81)$$

Случайные числа x^{LN}, имеющие логарифмически нормальное распределение с параметрами m_x и σ_x, получают по формуле

$$x^{LN} = e^{x^N}. \qquad (82)$$

Случайные числа x^P, соответствующие распределению Пуассона, получают на основе алгоритма

$$x^P = k, \qquad (83)$$

где k — такое наименьшее целое число, что

$$\sum_{i=1}^{k+1} [-\frac{1}{\lambda} \ln(1 - x_i^R)] > 1. \qquad (84)$$

1.3. Моделирование реализации случайных процессов

Для выполнения анализа процесса функционирования технической системы при случайных внешних воздействиях возникает необходимость моделирования этих воздействий. Реализации функций внешних воздействий на ЭВМ представляются в виде случайных последовательностей (значений воздействий в дискретные моменты времени), отображающих дискретные случайные процессы с заданными вероятностными характеристиками.

При моделировании стационарного случайного воздействия с нормальным распределением достаточно сформировать случайную последовательность с заданной корреляционной функцией. В основу алгоритмов формирования таких процессов положено линейное преобразование стационарной последовательности $\mathrm{x}_{\mathrm{k}}^{\mathrm{N}}$ независимых случайных чисел, имеющих нормальное распределение, в последовательность q_k. При этом случайная последовательность $\mathrm{x}_{\mathrm{k}}^{\mathrm{N}}$ подается на вход дискретного линейного фильтра, формирующего на выходе дискретный случайный процесс с заданной корреляционной функцией.

Алгоритмы формирования дискретных реализаций случайных процессов задаются рекуррентными соотношениями:

$$q_k = \sum_{i=0}^{N} c_i x_{k-i}, \; k = 0,1,2,...; \tag{85}$$

$$q_k = \sum_{l=1}^{L} a_l x_{k-1} - \sum_{j=1}^{m} b_j q_{k-j}, k = 0,1,2,..., \tag{86}$$

где a_l, b_j, c_i — параметры алгоритмов, определяемые по корреляционной функции $R_q(\tau)$ формируемого дискретного случайного процесса q_k.

Начальные значения q_k при $k = 0$ в этих алгоритмах для простоты можно выбирать нулевыми. При этом начальный участок моделируемого процесса будет несколько искажен переходным процессом, по окончании которого последовательность q_k, $k = 0,1,2...$ после некоторого значения k становится стационарной.

Для получения коэффициентов a_l, b_j, c_i, входящих в выражения скользящего суммирования (85) и (86), применяются разложение в ряд Фурье функции спектральной плотности; решение системы нелинейных алгебраических уравнений, правая часть которой определяется исходной корреляционной функцией; метод факторизации и др.

Рассмотрим наиболее часто встречающиеся на практике корреляционные функции $R_q(\tau)$ и алгоритмы моделирования случайных процессов, основанные на преобразовании последовательности x_k^N независимых нормально распределенных чисел с математическим ожиданием $m_x = 0$ и дисперсией $\sigma_x^2 = 1$ в последовательность q_k, характеризуемую корреляционной функцией

$$R_q(\tau) = R_q(kh) = M[qlql + k],\ k = 0,1,2,...,$$

(87)

где h — шаг дискретизации независимой переменной t.

При статистическом анализе случайных процессов в технических системах описание характеристик внешних воздействий обычно дается в виде корреляционных функций $R_q(\tau)$ или спектральных плотностей $S_q(\omega)$. Эти функции получают путем статистической обработки результатов экспериментальных исследований технических систем в реальных условиях их функционирования. Получаемые в результате экспериментов графики корреляционных функций аппроксимируют некоторыми функциями. Наиболее часто

используют экспоненциальные и экспоненциально-косинусные функции:

$$R_q(\tau) = \sigma_q^2 e^{-\alpha|\tau|};$$

(88)

$$R_q(\tau) = \sigma_q^2 e^{-\alpha|\tau|} \cos\beta\tau;$$

(89)

где σ_q^2 — дисперсия возмущающего воздействия *q(t)*, α — коэффициент, характеризующий затухание корреляционной функции; β — коэффициент, характеризующий колебательный процесс.

Разделив корреляционную функцию $R_q(\tau)$ на σ_q^2, получим нормированную корреляционную функцию $\rho_q(\tau)$. Графики нормированной экспоненциальной и экспоненциально-косинусной корреляционных функций показаны на рис. 12. Корреляционные функции $R_q(\tau)$ и $\rho_q(\tau)$ — четные. При $\tau = 0$ получаем $R_q(0) = \sigma_q^2$ и $p_q(0) = 1$. При $\tau > 0$ любое значение $R_q(\tau)$ меньше дисперсии случайного процесса σ_q^2, а $\rho_q(\tau)$ меньше единицы.

Корреляционная функция, определяемая выражением (89), относится к случайному процессу, содержащему периодическую составляющую.

Спектральная плотность стационарного случайного процесса $S_q(\omega)$ представляет собой функцию круговой частоты ω, которая равна преобразованию Фурье ковариационной функции $K(\tau)$ этого процесса:

$$S_q(\omega) = \int_{-\infty}^{\infty} K(\tau) e^{-j\omega\tau} d\tau. \qquad (90)$$

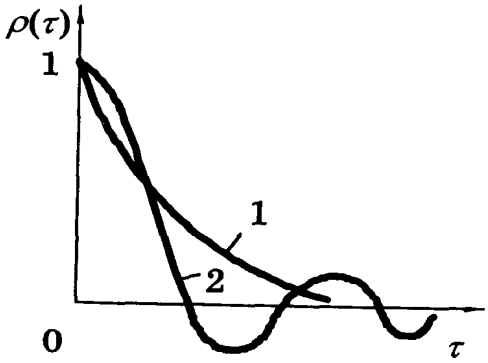


Рис. 12. Графики нормированных корреляционных функций: 1-экспоненциальной; 2- экспоненциально-косинусной

После преобразования выражения (90) для центрированного случайного процесса получим известное соотношение для области положительных частот

$$S_q(\omega) = \frac{2}{\pi}\int_0^\infty R_q(\tau)\cos\omega\,\tau\,d\tau. \qquad (91)$$

Для корреляционной функции $R_q(\tau)$ справедливо выражение

$$R_q(\tau) = \int_0^\infty S(\omega)\cos\omega\,\tau\,d\omega. \qquad (92)$$

Поскольку при $\tau = 0$ $R_q(0) = \sigma_q$, то на основании выражения (92) получаем

$$\sigma_q^2 = \int_0^\infty S(\omega)d\omega. \qquad (93)$$

Следовательно, площадь, ограниченная графиком функции $S(\omega)$ и осью частот ω , представляет собой дисперсию стационарного случайного процесса.

Перейдем к описанию алгоритмов формирования реализации дискретного случайного процесса с корреляционными функциями (88) и (89).

Последовательность ординат случайного процесса с корреляционной функцией (88) получают по формуле

$$q_k = a_0 x_k^N + b_1 q_{k-1},\ k = 0,1,2,\ldots,$$

(94)

где $a_0 = \sigma_q \sqrt{1 - b_1^2};\ b_1 = e^{-\alpha h};$ h — шаг дискретизации независимой переменной t; x_k^N — значения нормально распределенной случайной величины X^N с параметрами $m_x = 0$ и $\sigma_x = 1$.

Последовательность ординат случайного процесса с корреляционной функцией (89) получают по формуле

$$q_k = a_0 x_k^N + a_1 x_{k-1}^N + b_1 q_{k-1} + b_2 q_{k-2},\ k = 0,1,2,\ldots,$$

(95)

где

$$a_0 = \sigma_q b_0;\ a_1 = \sigma_q a_0 / b_0;\ b_0 = \sqrt{(c_1 + \sqrt{c_1^2 - 4c_0^2})/2};$$

$$b_1 = 2e^{-\alpha h}\cos\beta h;\ b_2 = e^{-2\alpha h};$$

$$c_0 = e^{-\alpha h}(e^{-2\alpha h} - 1)\cos\beta h;\ c_1 = 1 - e^{-4\alpha h}.$$

Если корреляционная функция случайного процесса представляет собой сумму выражений (88) и (89), то значение q_k равно сумме ординат, вычисленных по формулам (94) и (95). При этом x_k^N в этих формулах должны быть независимыми последовательностями нормально распределенных величин с параметрами $m_x = 0$ и $\sigma_x = 1$.

2 ЭКСПЕРИМЕНТАЛЬНЫЕ ФАКТОРНЫЕ МАТЕМАТИЧЕСКИЕ МОДЕЛИ

2.1. Особенности экспериментальных факторных моделей

Наряду с теоретическими математическими моделями при функциональном проектировании технических систем широко применяются экспериментальные факторные математические модели.

Теоретические модели имеют то преимущество, что они непосредственно описывают физические свойства технической системы. Коэффициенты уравнений теоретических моделей представляют собой параметры элементов технической системы (внутренние параметры системы) или некоторые комбинации этих параметров, а зависимые переменные — фазовые координаты системы. Они позволяют осуществлять имитационное моделирование процессов функционирования технической системы во времени, детально изучать изменение фазовых координат в зависимости от внешних воздействий (возмущающих и управляющих), анализировать устойчивость системы, качество переходных процессов, эффективность функционирования в условиях случайных внешних воздействий, близких к реальным, т. е. оценивать ее функциональную работоспособность и выполнение технических требований к системе.

Но функциональные теоретические модели сложных технических объектов представляют собой системы нелинейных дифференциальных уравнений высокого порядка (обычно не ниже 30-го порядка). Однократное решение такой системы уравнений на самых современных ЭВМ требует значительной затраты машинного времени (десятки и даже сотни минут). Следует при этом учитывать, что задачи проектирования носят ярко выраженный оптимизационный характер. Целью функционального проектирования является

выбор структуры на основе некоторого множества вариантов и определение оптимальных параметров технического объекта. Процедуры выбора структуры и оптимизационные алгоритмы требуют выполнения множества итераций, количество которых может достигать чисел второго и третьего порядков, причем на каждой итерации решается исходная система дифференциальных уравнений. Поэтому решение одной проектной задачи характеризуется огромными затратами машинного времени. Этим объясняется медленное внедрение методов функционального проектирования в конструкторских организациях. Вместе с тем без выполнения работ по функциональному проектированию невозможно обеспечить высокий технический уровень и конкурентоспособность создаваемых сложных технических объектов.

Затраты машинного времени можно значительно сократить, если на этапе оптимизации параметров использовать экспериментальную факторную математическую модель. *Экспериментальные факторные модели,* в отличие от теоретических, не используют физических законов, описывающих происходящие в объектах процессы, а представляют собой некоторые формальные зависимости выходных параметров от внутренних и внешних параметров объектов проектирования.

Экспериментальная факторная модель может быть построена на основе проведения экспериментов непосредственно на самом техническом объекте *(физические эксперименты)* либо *вычислительных экспериментов* на ЭВМ с теоретической моделью. При создании новых технических объектов физический эксперимент проводится на прототипах или аналогах, а иногда на макетных образцах. Однако физические эксперименты требуют огромных затрат материальных и временных ресурсов, поэтому их выполняют обычно в тех случаях, когда возникает необходимость поиска путей совершенствования существующих технических систем,

когда сложность этих систем и условий их функционирования не позволяет надеяться на требуемую точность их математического описания теоретическими методами.

При функциональном проектировании факторные модели наиболее часто получают на основе вычислительных экспериментов на ЭВМ с теоретической моделью.

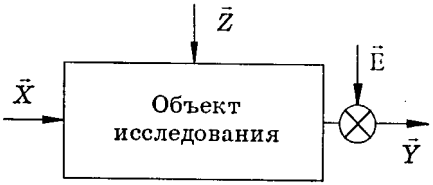


Рис. 13. Схема объекта исследования при построении экспериментальной факторной модели

При построении экспериментальной факторной модели объект моделирования (проектируемая техническая система) представляется в виде «черного ящика», на вход которого подаются некоторые переменные $\vec{X}$ и $\vec{Z}$, а на выходе можно наблюдать и регистрировать переменные $\vec{Y}$ (рис. 13). В число входных переменных $\vec{X}$ и $\vec{Z}$ входят внутренние и внешние параметры объекта проектирования, подлежащие оптимизации, а выходными переменными «черного ящика» являются выходные параметры объекта, характеризующие его эффективность и качество процессов функционирования, выбираемые в качестве критериев оптимальности.

В процессе проведения эксперимента изменение переменных $\vec{X}$ и $\vec{Z}$ приводит к изменениям выходных переменных $\vec{Y}$. Для построения факторной модели необходимо регистрировать эти изменения и осуществить необходимую их статистическую обработку для определения параметров модели.

При проведении физического эксперимента переменными $\vec{X}$ можно управлять, изменяя их величину по заданному закону. Переменные $\vec{Z}$ — неуправляемые, принимающие случайные значения. При этом значения переменных $\vec{X}$ и $\vec{Z}$ можно контролировать и регистрировать с помощью соответствующих измерительных приборов. Кроме того, на объект воздействуют некоторые переменные $\vec{E}$, которые нельзя наблюдать и контролировать. Переменные $\vec{X} = (x_1, x_2, \ldots, x_n)$ называют *контролируемыми и управляемыми;* переменные $\vec{Z} = (z_1, z_2, \ldots, z_m)$ — *контролируемыми, но неуправляемыми,* а переменные $\vec{E} = (e_1, e_2, \ldots, e_l)$ — *неконтролируемыми и неуправляемыми.*

Переменные $\vec{X}$ и $\vec{Z}$ называют *факторами.* Факторы $\vec{X}$ являются управляемыми и изменяются как *детерминированные переменные,* а факторы $\vec{Z}$ неуправляемые, изменяемые во времени случайным образом, т. е. $\vec{Z}$ представляют собой *случайные процессы.* Пространство контролируемых переменных — факторов $\vec{X}$ и $\vec{Z}$ — образует *факторное пространство.*

Выходная переменная $\vec{Y}$ представляет собой вектор зависимых переменных моделируемого объекта. Ее называют *откликом,* а зависимость $\vec{Y}$ от факторов $\vec{X}$ и $\vec{Z}$ — *функцией отклика.* Геометрическое представление функции отклика называют *поверхностью отклика.*

Переменная $\vec{E}$ действует в процессе эксперимента бесконтрольно. Если предположить, что факторы $\vec{X}$ и $\vec{Z}$ стабилизированы во времени и сохраняют постоянные значения, то под влиянием переменных $\vec{E}$ функция отклика $\vec{Y}$ может меняться как систематическим, так и случайным образом. В первом случае говорят о *систематической*

помехе, а во втором — о *случайной помехе.* При этом полагают, что случайная помеха обладает вероятностными свойствами, не изменяемыми во времени.

Возникновение помех обусловлено ошибками методик проведения физических экспериментов, ошибками измерительных приборов, неконтролируемыми изменениями параметров и характеристик объекта и внешней среды, включая воздействия тех переменных, которые в принципе могли бы контролироваться экспериментатором, но не включены им в число исследуемых факторов (вследствие трудностей их измерения, по ошибке или незнанию). Помехи могут быть также обусловлены неточностью физического или математического моделирования объектов.

В вычислительных экспериментах объектом исследования является теоретическая математическая модель, на основе которой необходимо получить экспериментальную факторную модель. Для ее получения необходимо определить структуру и численные значения параметров модели.

Под *структурой модели* понимается вид математических соотношений между факторами $\vec{X}$, $\vec{Z}$ и откликом $\vec{Y}$. *Параметры* представляют собой коэффициенты уравнений факторной модели. Структуру модели обычно выбирают на основе априорной информации об объекте с учетом назначения и последующего использования модели. Задача определения параметров модели полностью формализована. Она решается методами *регрессионного анализа. Экспериментальные факторные модели* называют также *регрессионными моделями.*

Регрессионную модель можно представить выражением

$$\vec{Y} = \vec{\varphi}(\vec{X}, \vec{B}),$$

где $\vec{B}$ — вектор параметров факторной модели.

Вид вектор-функции $\vec{\varphi}$ определяется выбранной структурой модели и при выполнении регрессионного анализа считается заданным, а параметры $\vec{B}$ подлежат определению на основе

результатов эксперимента, проводимого в условиях действия помехи $\vec{E}$, представляемой в виде аддитивной составляющей функции отклика $\vec{Y}$ (рис. 13).

Эксперимент — это система операций, воздействий и (или) наблюдений, направленных на получение информации об объекте при исследовательских испытаниях.

Опыт — воспроизведение исследуемого явления в определенных условиях проведения эксперимента при возможности регистрации его результатов. Опыт — отдельная элементарная часть эксперимента.

Различают эксперименты пассивные и активные.

Пассивным называется такой эксперимент, когда значениями факторов управлять нельзя, и они принимают случайные значения. Это характерно для многих технических объектов при проведении на них физических экспериментов. В таком эксперименте существуют только факторы $\vec{Z}$. В процессе эксперимента в определенные моменты времени измеряются значения факторов $\vec{Z}$ и функций откликов $\vec{Y}$. После проведения N опытов полученная информация обрабатывается статистическими методами, позволяющими определить параметры факторной модели. Такой подход к построению математической модели лежит в основе *метода статистических испытаний (Монте-Карло).*

Активным называется такой эксперимент, когда значениями факторов задаются и поддерживаются неизменными на заданных уровнях в каждом опыте в соответствии с планом эксперимента. Следовательно, в этом случае существуют только управляемые факторы $\vec{X}$. Однако в связи с тем, что в активном эксперименте также действует аддитивная помеха $\vec{E}$, реализации функций отклика $\vec{Y}$ представляют собой случайные величины, несмотря на то, что варьируемые факторы $\vec{X}$ детерминированы. Поэтому здесь также, как и в пассивном эксперименте, построение экспериментальной

факторной модели требует статистической обработки получаемых результатов опытов.
Основные особенности экспериментальных факторных моделей следующие: они статистические; представляют собой сравнительно простые функциональные зависимости между оценками математических ожиданий выходных параметров объекта от его внутренних и внешних параметров; дают адекватное описание установленных зависимостей лишь в области факторного пространства, в которой реализован эксперимент. Статистическая регрессионная модель описывает поведение объекта в среднем, характеризуя его неслучайные свойства, которые в полной мере проявляются лишь при многократном повторении опытов в неизменных условиях.

2.1.1. Основные принципы планирования эксперимента

Для получения адекватной математической модели необходимо обеспечить выполнение определенных условий проведения эксперимента. Модель называют *адекватной,* если в оговоренной области варьирования факторов $\vec{X}$ полученные с помощью модели значения функций отклика $\vec{Y}$ отличаются от истинных не более чем на заданную величину.
Методы построения экспериментальных факторных моделей рассматриваются в *теории планирования эксперимента.*
Цель планирования эксперимента — получение максимума информации о свойствах исследуемого объекта при минимуме опытов. Такой подход обусловлен высокой стоимостью экспериментов, как физических, так и вычислительных, и вместе с тем необходимостью построения адекватной модели.
Планирование осуществляют как активного, так и пассивного эксперимента. Планируемый активный эксперимент при прочих равных условиях точнее и информативнее, а иногда и

дешевле пассивного. Это следует учитывать при выборе вида эксперимента. В вычислительном эксперименте, в отличие от физического, нет никаких ограничений на выбор управляемых факторов и характер их изменения. Поэтому вычислительные эксперименты обычно всегда реализуются как активные. В дальнейшем будут рассматриваться в основном вопросы, связанные с планированием активных экспериментов.

При планировании активных экспериментов используются следующие принципы:

отказ от полного перебора всех возможных состояний объекта;

постепенное усложнение структуры математической модели;

сопоставление результатов эксперимента с величиной случайных помех;

рандомизация опытов;

оптимальное планирование эксперимента.

Детальное представление о свойствах поверхности отклика может быть получено лишь при условии использования густой дискретной сетки значений факторов, покрывающей все факторное пространство. В узлах этой многомерной сетки находятся точки плана, в которых проводятся опыты. В этом случае в принципе можно получить факторную модель, которая будет практически почти полностью соответствовать исходной теоретической модели. Однако в большинстве случаев при решении практических задач, для которых используется факторная модель, такого детального описания не требуется. Выбор структуры факторной модели основан на постулировании определенной степени гладкости поверхности отклика. Поэтому с целью уменьшения количества опытов принимают небольшое число точек плана, для которых осуществляется реализация эксперимента.

В отсутствие априорной информации о свойствах функции отклика нет смысла сразу строить сложную математическую

модель объекта. Если проверка этой модели на адекватность не дает удовлетворительного результата, ее постепенно усложняют путем изменения структуры (например, повышая степень полинома, принятого в качестве факторной модели, или вводя в модель дополнительные факторы и т. п.). При этом используются результаты опытов, выполненных при построении простой модели, и проводится некоторое количество дополнительных опытов.

При большом уровне случайной помехи получается большой разброс значений функции отклика $\vec{Y}$ в опытах, проведенных в одной и той же точке плана. В этом случае оказывается, что чем выше уровень помехи, тем с большей вероятностью простая модель окажется работоспособной. Чем меньше уровень помехи, тем точнее должна быть факторная модель.

Кроме случайной помехи при проведении эксперимента может иметь место систематическая помеха. Наличие этой помехи практически никак не обнаруживается, и результат ее воздействия на функцию не поддается контролю. Однако, если путем соответствующей организации проведения опытов искусственно создать случайную ситуацию, то систематическую помеху можно перевести в разряд случайных. Такой принцип организации эксперимента называют *рандомизацией* систематически действующих помех.

Наличие помех приводит к ошибкам эксперимента. *Ошибки* подразделяют на *систематические* и *случайные,* соответственно наименованиям вызывающих их факторов — помех.

В вычислительных активных экспериментах ошибки характерны только для определяемых значений функций отклика. Если исходить из целей построения факторных моделей на основе теоретических моделей, полагая, что теоретические модели дают точное описание физических свойств технического объекта, а регрессионная модель

является ее аппроксимацией, то значения функций отклика будут содержать только случайную ошибку. В этом случае необходимости в рандомизации опытов не возникает.

Рандомизацию опытов осуществляют только в физических экспериментах. Следует отметить, что в этих экспериментах систематическую ошибку может порождать наряду с отмеченными в предыдущем параграфе факторами также неточное задание значений управляемых факторов, обусловленное некачественной калибровкой приборов для их измерения (инструментальная ошибка), конструктивными или технологическими факторами.

К факторам в активном эксперименте предъявляются определенные требования. Они должны быть:

1) *управляемыми* (установка заданных значений и поддержание постоянными в процессе опыта);

2) *совместными* (их взаимное влияние не должно нарушать процесс функционирования объекта);

3) *независимыми* (уровень любого фактора должен устанавливаться независимо от уровней остальных);

4) *однозначными* (одни факторы не должны быть функцией других);

5) *непосредственно влияющими на выходные параметры.*

В вычислительном эксперименте реализация трех первых требований не создает никаких затруднений, а в физическом эксперименте могут возникнуть сложности и даже невозможность их осуществления, что приведет к необходимости замены активного эксперимента пассивным.

Функции отклика должны быть:

1) *численно измеряемыми;*

2) иметь четкий физический смысл;

3) *однозначными* (характеризовать только одно свойство объекта);

4) *информативными* (полностью характеризовать определенное свойство объекта);

5) *статистически эффективными* (измеряться с достаточной точностью с целью сокращения дублирования опытов).

2.1.2 План эксперимента

При проведении активного эксперимента задается определенный план варьирования факторов, т. е. эксперимент заранее планируется.

План эксперимента — совокупность данных, определяющих число, условия и порядок реализации опытов.

Планирование эксперимента — выбор плана эксперимента, удовлетворяющего заданным требованиям.

Точка плана — упорядоченная совокупность численных значений факторов, соответствующая условиям проведения опыта, т. е. точка факторного пространства, в которой проводится эксперимент. Точке плана с номером i соответствует вектор-строка

$$\vec{X}_i = (X_{i1}, X_{i2}, ..., X_{in}).$$

Общая совокупность таких векторов $\vec{X}_i$, $i = \overline{1, L}$, , образует план эксперимента, а совокупность различных векторов, число которых обозначим N, — *спектр плана*.

В активном эксперименте факторы могут принимать только фиксированные значения. Фиксированное значение фактора называют уровнем фактора. Количество принимаемых уровней факторов зависит от выбранной структуры факторной модели и принятого плана эксперимента. Минимальный $X_{j\ min}$ и максимальный $X_{j\ max}$, $\mathrm{j} = \overline{1, \mathrm{n}}$ (n — число факторов), уровни всех факторов выделяют в факторном пространстве некоторый гиперпараллелепипед, представляющий собой *область планирования*. В области планирования находятся все возможные значения факторов, используемые в эксперименте.

Вектор $\vec{X}^0 = (X_1^0, X_2^0, ..., X_n^0)$ задает точку центра области планирования. Координаты этой точки X_j^0 обычно выбирают из соотношения

$$X_j^0 = (X_{j\max} + X_{j\min})/2. \qquad (96)$$

Точку $\vec{X}^0$ называют *центром эксперимента.* Она определяет основной уровень факторов $\vec{X}^0$, $j = \overline{1,n}$. Центр эксперимента стремятся выбрать как можно ближе к точке, которая соответствует искомым оптимальным значениям факторов. Для этого используется априорная информация об объекте.

Интервалом (или *шагом) варьирования фактора* X_j называют величину, вычисляемую по формуле

$$\Delta X_j = (X_{j\max} - X_{j\min})/2, \quad j = \overline{1,n}. \qquad (97)$$

Факторы нормируют, а их уровни кодируют. В кодированном виде верхний уровень обозначают *+1*, нижний *-1*, а основной *0*. Нормирование факторов осуществляют на основе соотношения

$$x_j = (X_j - X_j^0)/\Delta X_j, \quad j = \overline{1,n}. \qquad (98)$$

Для переменных x_j начало координат совмещено с центром эксперимента, а в качестве единиц измерения используются интервалы варьирования факторов. Геометрическое представление области планирования при двух факторах показано на рис. 14. Центр эксперимента находится в точке *0* с координатами X_1^0, X_2^0. Точки *1,2,3,4* являются точками плана эксперимента. Например, значения факторов X_1 и X_2 в точке *1* равны соответственно $-X_{1min}$, X_{2min}, а нормированные их значения $x_{1min} = -1$, $x_{2min} = -1$.

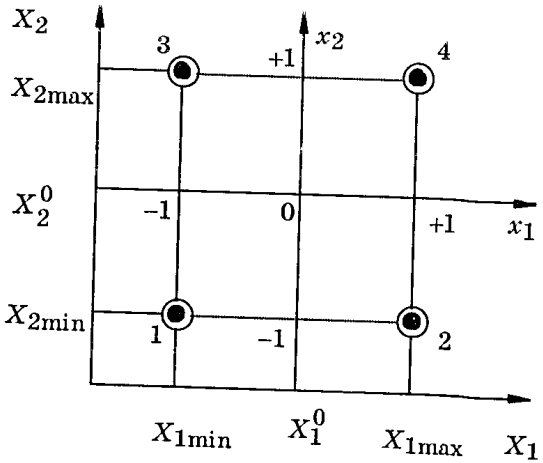


Рис. 14. Геометрическое представление области планирования при двух факторах X_1 и X_2

В дальнейшем будем предполагать, что в планах активных экспериментов факторы нормированы.

План эксперимента удобно представлять в матричной форме. План эксперимента задается либо матрицей плана, либо матрицей спектра плана в совокупности с матрицей дублирования.

Матрица плана представляет собой прямоугольную таблицу, содержащую информацию о количестве и условиях проведения опытов. Строки матрицы плана соответствуют опытам, а столбцы — факторам. Размерность матрицы плана $L \times n$, где L — число опытов, n — число факторов. При проведении повторных (дублирующих) опытов в одних и тех же точках плана матрица плана содержит ряд совпадающих строк.

Матрица спектра плана — это матрица, в которую входят только различающиеся между собой строки матрицы плана. Размерность матрицы спектра плана $N \times n$, где N — число точек плана, различающихся между собой хотя бы одной координатой X_{ij}, $\mathrm{i} = \overline{1, \mathrm{N}}$, $\mathrm{j} = \overline{1, \mathrm{n}}$.

Матрица спектра плана имеет вид

$$X = \begin{bmatrix} \vec{X}_1 \\ \vec{X}_2 \\ \vdots \\ \vec{X}_i \\ \vdots \\ \vec{X}_N \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_{N1} & X_{N2} & \dots & X_{Nj} & \dots & X_{Nn} \end{bmatrix}, \quad (99)$$

где $\vec{X}_i$ — вектор, определяющий нормированные значения координат точки плана в *i*-ом опыте; X_{ij} — нормированное значение *j*-го фактора в *i*-ом опыте.

Матрица дублирования — квадратная диагональная матрица m, диагональные элементы которой равны числам параллельных опытов в соответствующих точках спектра плана:

$$m = \begin{bmatrix} m_1 & 0 & \dots & 0 \\ 0 & m_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & m_N \end{bmatrix}. \quad (100)$$

Опыты при выполнении эксперимента проводятся в последовательности, предусмотренной матрицей плана. Эта матрица составляется лишь при необходимости рандомизации опытов, когда в результатах эксперимента можно ожидать наличие систематических ошибок. Для выбора случайной последовательности опытов используется таблица равномерно распределенных случайных чисел. Первое число таблицы выбирают произвольно, желательно случайным образом, а затем, начиная с этого числа, выписывают *L* чисел таблицы, где *L* — число опытов (с учетом их дублирования). При этом числа, большие *L*, а также уже выписанные, отбрасываются.

В вычислительных экспериментах опыты проводят в соответствии с матрицей спектра плана, так как предполагается отсутствие систематических ошибок и поэтому нет необходимости в рандомизации опытов.

2.2. Регрессионный анализ

Регрессионный анализ проводится с целью получения по экспериментальным данным регрессионных моделей, представляющих собой экспериментальные факторные модели. Задачей регрессионного анализа является определение параметров экспериментальных факторных моделей объектов проектирования или исследования, т. е. определение коэффициентов уравнений моделей при выбранной их структуре.

Регрессионный анализ включает три основных этапа:

1) статистический анализ результатов эксперимента;

2) получение оценок $\vec{b}$ искомых коэффициентов регрессии $\vec{\beta}$;

3) оценку адекватности и работоспособности полученной экспериментальной факторной модели технической системы.

Под структурой экспериментальной факторной математической модели понимается вид математических соотношений между факторами $\vec{X}$, $\vec{Z}$ и откликом $\vec{Y}$. В качестве факторов принимают внутренние и внешние параметры технической системы, подлежащие оптимизации в процессе ее проектирования. Внутренние параметры системы — это параметры ее элементов, внешние — это параметры внешней среды, в условиях воздействий которой осуществляется функционирование системы. Функциями отклика $\vec{Y}$ являются выходные параметры технической системы, характеризующие ее эффективность и качество процессов функционирования. Выходные параметры системы принимаются в качестве критериев оптимальности.

Как уже отмечалось, структура факторной модели выбирается на основе априорной информации, используя принцип постепенного ее усложнения. Параметры факторной

математической модели определяются методами регрессионного анализа. При определении параметров этими методами нет необходимости различать виды факторов, т. е. подразделять факторы на управляемые $\vec{X}$ и неуправляемые $\vec{Z}$. Поэтому в дальнейшем все они будут обозначаться буквой $\vec{X}$. Тогда факторную модель можно представить векторным уравнением регрессии вида

$$\vec{Y} = \vec{\varphi}(\vec{X}, \vec{B}). \qquad (101)$$

Определение параметров $\vec{B}$ этой модели будем рассматривать на примере одного уравнения $Y = \vec{\varphi}(\vec{X}, \vec{B})$. Для определения параметров используются результаты эксперимента. Результаты эксперимента можно представить функцией вида

$$Y = \varphi(\vec{X}) + \varepsilon, \qquad (102)$$

где ε — аддитивная помеха случайного характера с нормальным законом распределения.

Так как каждый опыт проводится при определенном сочетании уровней факторов $\vec{X}$, то функцию $\varphi(\vec{X})$ представим выражением

$$\varphi(\vec{X}) = \sum_{j=0}^{d} \beta_j f_j(\vec{X}), \qquad (103)$$

где β_j — j-ый элемент вектора искомых коэффициентов уравнения регрессии: $\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_d^T)$, $f_j(\vec{X})$ — j-ая базисная функция — элемент вектора базисных функций $\vec{f}(\vec{X}) = [f_0(\vec{X}), f_1(\vec{X}), \ldots, f_d(\vec{X})]^T$.

В качестве базисных функций используют переменные простейших полиномов, системы ортогональных полиномов (Эрмита, Лежандра, Лаггера и др.), тригонометрические функции. Наиболее часто пользуются простейшими полиномами первой и второй степеней. Например, полином

первой степени, описывающий функцию отклика y при двух факторах x_1 и x_2, может иметь вид

$$y = b_0 + b_1 x_1 + b_2 x_2, \qquad (104)$$

или

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2, \qquad (105)$$

а полином второй степени

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 + b_4 x_1^2 + b_5 x_2^2. \qquad (106)$$

Базисные функции в случае использования последнего выражения имеют вид:

$$f_0(\vec{X}) = 1;\ f_1(\vec{X}) = x_1;\ f_2(\vec{X}) = x_2;\ f_3(\vec{X}) = x_1 x_2;\ f_4(\vec{X}) = x_1^2;$$

$$f_5 = (\vec{X}) = x_2^2.$$

Если уравнение регрессии имеет вид выражений вида (104), (105), его называют уравнением *линейной регрессии* (линейной регрессией или регрессией первого порядка), а если содержит факторы во второй и более высокой степени — *нелинейной регрессией* (регрессией соответствующего порядка).

Линейная регрессия может представлять как линейную математическую модель, так и нелинейную, в зависимости от того, содержит ли она *линейные эффекты* (как в выражении (104) или наряду с ними также *эффекты взаимодействия* (как в выражении (105)). Линейным называют эффект, характеризующий линейную зависимость выходного параметра *y* от соответствующего фактора x_i. Эффектом взаимодействия называют эффект, характеризующий совместное влияние нескольких факторов на y (например, в выражении (105) x_1x_2). Эффекты взаимодействия двух факторов называют парным взаимодействием, трех факторов — тройным взаимодействием и т.д.

Как всякий статистический метод, регрессионный анализ применим при определенных предпосылках (постулатах).

1. Аддитивная помеха ε — случайная нормально распределенная величина с параметрами $m_\varepsilon = 0$ и $\sigma_\varepsilon^2 = const$. В

этом случае функция отклика Y также случайная величина с нормальным законом распределения. Гипотезу о нормальном распределении Y можно проверить по критерию Пирсона.

2. Постоянство дисперсии помехи означает, что интенсивность ошибки определения Y не меняется при изменении уровня факторов в процессе эксперимента. Выполнение этого постулата проверяется по критерию однородности дисперсии в разных точках опыта.

3. Значения факторов в активном эксперименте — неслучайные величины. Это означает, что установление каждого фактора на заданном уровне и удерживание его на этом уровне во время опыта точнее, чем ошибка воспроизводимости. В вычислительном эксперименте это выполняется однозначно, а в физическом вклад, вносимый ошибками измерения факторов $\vec{X}$, должен быть пренебрежимо малым в сравнении с действием других неконтролируемых факторов, образующих ошибку ε определения функции Y.

4. Значения помехи ε в различных точках опыта некоррелированы. Для обеспечения этого требования используется рандомизация опытов.

В пассивном эксперименте условие некоррелированности помехи обеспечивают путем соответствующего выбора временного интервала съема информации об условиях и результатах опытов.

5. Векторы-столбцы базисных функций должны быть линейно независимыми. Выполнение этого требования необходимо для получения раздельных оценок $\vec{b}$ всех коэффициентов регрессии $\vec{\beta}$. В активном эксперименте оно обеспечивается соответствующим выбором спектра плана эксперимента. При этом число опытов N (без учета дублирования) должно быть не меньше, чем число оцениваемых коэффициентов $N_в$,

т. е. $N \geq N_в$.

В пассивном эксперименте линейная зависимость между столбцами практически исключается, так как факторы неуправляемы и принимают случайные значения в разных опытах, но может наблюдаться сильная коррелированность столбцов, что повлечет за собой большие ошибки вычисления коэффициентов регрессии. Для выявления коррелированности столбцов проводят корреляционный анализ результатов пассивного эксперимента.

2.2.1. Оценка параметров регрессионной модели

Исходными данными для получения оценок параметров регрессионной модели технической системы (т. е. оценок $\vec{b}$ искомых коэффициентов регрессии $\vec{\beta}$) является информация о значения управляемых факторов $\vec{X}$ (или неуправляемых — при проведении пассивного эксперимента) и функции отклика *Y*. Эту информацию можно представить в виде матрицы *X* значений факторов во всех *N* опытах, предусмотренных спектром плана эксперимента, и вектора-столбца $\vec{Y}$ полученных в этих опытах значений функции отклика *Y*:

$$X = \begin{bmatrix} \vec{X}_1 \\ \vec{X}_2 \\ \vec{X}_i \\ \vec{X}_N \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2n} \\ X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{in} \\ X_{N1} & X_{N2} & \dots & X_{Nj} & \dots & X_{Nn} \end{bmatrix}; \quad (107)$$

$$\vec{Y} = (y_1, y_2, ..., y_i, ..., y_N)^T, \quad (108)$$

где $\vec{X}_i = (X_{i1}, X_{i2}, ..., X_{in})$ — вектор-строка значений факторов в 1-м опыте; X_{ij} — значение *j*-го фактора в *i*-м

опыте; n — количество факторов; N — количество опытов; y_i — значение функции отклика Y в i-ом опыте (если проводились параллельные опыты, т. е. осуществлялось дублирование опытов, то вместо y_i используются оценки их математических ожиданий, т. е. выборочные средние $\vec{y}_i$).

Значения базисных функций во всех опытах представляют собой матрицу F, называемую *матрицей базисных функций:*

$$F = \begin{bmatrix} \vec{f}_1(\vec{X}_1) \\ \vec{f}_2(\vec{X}_2) \\ \vec{f}_i(\vec{X}_i) \\ \vec{f}_N(\vec{X}_N) \end{bmatrix} = \begin{bmatrix} f_{10} & f_{11} & f_{12} & \cdots & f_{1k} & \cdots & f_{1d} \\ f_{20} & f_{21} & f_{22} & \cdots & f_{2k} & \cdots & f_{2d} \\ f_{i0} & f_{i1} & f_{i2} & \cdots & f_{ik} & \cdots & f_{id} \\ f_{N0} & f_{N1} & f_{N2} & \cdots & f_{Nk} & \cdots & f_{Nd} \end{bmatrix}, \quad (109)$$

где f_{ik} — значение k-ой базисной функции в i-ом опыте; $\vec{f}_i(\vec{X}_i) = (f_{i0}, f_{i1}, f_{i2},..., f_{ik},..., f_{id})$ – вектор-строка значений базисных функций в i-ом опыте.

Используя информацию об X, $\vec{Y}$ и F, необходимо найти оценки коэффициентов регрессии, представляемые вектором-столбцом

$$\vec{b}^T = (b_0, b_1, b_2,..., b_k,..., b_d), \quad (110)$$

где b_k — значение оценки коэффициента регрессии при базисной функции $f_k(\vec{X})$.

Так как функция отклика Y — случайная величина, поскольку на ее значения в различных опытах оказывает влияние случайная помеха ε, то оценки коэффициентов регрессии будут случайными величинами.

Уравнение регрессии устанавливает зависимость между оценкой математического ожидания функции отклика $\bar{y}$ и факторами $\vec{X} = (x_1, x_2,..., x_n)$. Общий вид этой зависимости

$$\bar{y} = \sum_{k=0}^{d} b_k f_k(\vec{X}). \quad (111)$$

В связи с наличием помехи значение функции отклика в i-м опыте y_i будет отличаться от $\overline{y}_i$. Для определения y_i можно составить выражение

$$y_i = b_0 f_{i0} + b_1 f_{i1} + ... + b_k f_{ik} + ... + b_d f_{id} + \varepsilon_i, \; i = \overline{1, N}, \quad (112)$$

где ε_i — невязка уравнения регрессии в i-м опыте.

Невязка характеризует отклонение значений функции отклика в опытах от получаемых с помощью регрессионной модели (111). Она возникает по двум причинам: из-за ошибки эксперимента и из-за непригодности (приближенности) выбранной структуры факторной математической модели. Причем, эти причины смешаны и нельзя сказать, какая из них преобладает.

Если постулировать, что модель пригодна, то невязка будет порождаться только ошибкой опыта. Тогда для определения коэффициентов уравнения (111) невязку надо минимизировать. Для этого в регрессионном анализе используется *метод наименьших квадратов* (МНК). Составляется функция, представляющая собой сумму квадратов невязок, и осуществляется ее минимизация, т.е.

$$E = \sum_{i=1}^{N} \varepsilon_i^2 \to \min. \quad (113)$$

Подставим значение ε_i из выражения (112):

$$E = \sum_{i=1}^{N} \left[y_i - (b_0 f_{i0} + b_1 f_{i1} + ... + b_k f_{ik} + ... + b_d f_{id}) \right]^2 \to \min. \quad (114)$$

В выражении (114) коэффициенты b_k рассматриваются как неизвестные переменные, которые наилучшим образом соответствуют полученным результатам эксперимента. Значения этих коэффициентов, при которых достигается

минимум функции E, имеет место при равенстве нулю частных производных этой функции по переменным b_0, b_1,..., b_d:

$$\frac{\partial E}{\partial b_0} = -2\sum_{i=1}^{N}\left[y_i - (b_0 f_{i0} + b_1 f_{i1} + ... + b_d f_{id})\right] f_{i0} = 0;$$
$$\frac{\partial E}{\partial b_1} = -2\sum_{i=1}^{N}\left[y_i - (b_0 f_{i0} + b_1 f_{i1} + ... + b_d f_{id})\right] f_{i1} = 0;$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\frac{\partial E}{\partial b_d} = -2\sum_{i=1}^{N}\left[y_i - (b_0 f_{i0} + b_1 f_{i1} + ... + b_d f_{id})\right] f_{id} = 0.$$

После преобразований получим систему линейных неоднородных алгебраических уравнений относительно искомых оценок коэффициентов регрессии b_0, b_1,..., b_d:

$$b_0\sum_{i=1}^{N} f_{i0}^2 + b_1\sum_{i=1}^{N} f_{i1} f_{i0} + ... + b_d\sum_{i=1}^{N} f_{id} f_{i0} = \sum_{i=0}^{N} y_i f_{i0};$$
$$b_0\sum_{i=1}^{N} f_{i0} f_{i1} + b_1\sum_{i=1}^{N} f_{i1}^2 + ... + b_d\sum_{i=1}^{N} f_{id} f_{i1} = \sum_{i=1}^{N} y_i f_{i1};$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$b_0\sum_{i=1}^{N} f_{i0} f_{id} + b_1\sum_{i=1}^{N} f_{i1} f_{id} + ... + b_d\sum_{i=1}^{N} f_{id}^2 = \sum_{i=1}^{N} y_i f_{id}.$$

(115)

Очевидно, что коэффициенты при неизвестных переменных $\vec{b}$ этой системы уравнений являются элементами матрицы Φ, определяемой из выражения

$$\Phi = F^T F, \qquad (116)$$

в котором F представляет собой матрицу базисных функций (109). Значения элементов матрицы F известны из проведенного эксперимента. Следовательно, элементы

матрицы Φ оказываются известными коэффициентами системы уравнений (115). Выпишем матрицу Φ:

$$\Phi = \begin{bmatrix} \sum_{i=1}^{N} f_{i0}^2 & \sum_{i=1}^{N} f_{i1} f_{i0} & \cdots & \sum_{i=1}^{N} f_{id} f_{i0} \\ \sum_{i=1}^{N} f_{i0} f_{i1} & \sum_{i=1}^{N} f_{i1}^2 & \cdots & \sum_{i=1}^{N} f_{id} f_{i1} \\ \sum_{i=1}^{N} f_{i0} f_{id} & \sum_{i=1}^{N} f_{i1} f_{id} & \cdots & \sum_{i=1}^{N} f_{id}^2 \end{bmatrix}. \quad (117)$$

Матрицу Φ называют *информационной матрицей Фишера*. Она содержит $(d+1)$ строк и $(d+1)$ столбцов, причем элемент j-ой строки k-го столбца представляет собой сумму $\sum_{i=1}^{N} f_{ij} f_{ik}$.

Матрица Φ симметрична относительно главной диагонали, что упрощает составление системы алгебраических уравнений (115) для регрессионной модели.

Систему уравнений (115) можно также записать в матричной форме:

$$\Phi \vec{B} = F^T \vec{Y}. \quad (118)$$

Система уравнений (115) имеет единственное решение, если определитель матрицы Φ не равен нулю. В этом случае матрица Φ будет не вырожденной. Выполнение пятой предпосылки регрессионного анализа, изложенной в предыдущем параграфе, исключает возникновение вырожденности.

Решение системы уравнений (115) обычно осуществляют методом Гаусса. При небольшом числе определяемых коэффициентов b_k можно использовать правило Крамера.

Полученные методом наименьших квадратов оценки $b_0, b_1,..., b_d$ действительных значений коэффициентов регрессии $\beta_0, \beta_1,..., \beta_d$ обладают следующими свойствами:

1) математические ожидания оценок

$M[b_j] = \beta_j$, $j = \overline{0,d}$, т. е. оценки b_j несмещенные;

2) дисперсии оценок коэффициентов регрессии минимальны и равны

$$\sigma^2_{b_j} = M\{(b_j - M[b_j])^2\} = M\{(b_j - \beta_j)^2\} = \sigma^2_\varepsilon C_{ij},$$

(119)

а корреляционный момент

$$\mu_{11}(b_j, b_k) = M\{(b_j - M[b_j])(b_k - M[b_k])\} =$$
$$= M\{(b_j - \beta_j)(b_k - \beta_k)\} = \sigma^2_\varepsilon C_{jk},$$

(120)

где C_{jj}, C_{jk} — элементы матрицы Φ^{-1}, обратной к информационной; σ^2_ε — дисперсия случайной помехи;

3) оценки $b_0, b_1, ..., b_d$ подчиняются совместному $(d + 1)$-мерному нормальному распределению.

2.3. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

2.3.1. Основные понятия

Корреляционный анализ, разработанный К. Пирсоном и Дж. Юлом, является одним из методов статистического анализа взаимозависимости нескольких признаков — компонент случайного вектора x.

Одним из основных показателей взаимозависимости двух случайных величин является **парный коэффициент корреляции,** *служащий мерой линейной статистической зависимости между этими величинами.* Следовательно, этот показатель соответствует своему прямому назначению, когда статистическая связь между соответствующими признаками в генеральной совокупности линейна. То же самое касается **частных и совокупных коэффициентов корреляции.** Одним из требований, определяющих корреляционный метод, является **требование линейности статистической**

связи, т. е. линейности всевозможных уравнений (средней квадратической) регрессии.

Указанные условия выполняются, если генеральная совокупность распределена по многомерному нормальному закону.

В настоящее время корреляционный анализ (корреляционная модель) определяется как метод, применяемый тогда, когда данные наблюдений или эксперимента можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

Основная задача корреляционного анализа состоит в оценке k(k+3)/2 параметров, определяющих нормальный закон распределения k-мерного вектора x, *в частности, корреляционной матрицы генеральной совокупности X, по выборке.*

Для значимых парных коэффициентов корреляции имеет смысл указать более предпочтительные точечные или интервальные оценки.

Далее следует оценить и проверить значимость множественных коэффициентов корреляции или детерминации всевозможных подсистем системы $x_j(j=1,k)$, содержащих три и более различных случайных величин x_j.

Для выяснения "чистых", истинных взаимозависимостей следует проанализировать выборочные частные коэффициенты корреляции.

Таким образом, основная задача позволяет определить расположение "облака" точек в пространстве *k* измерений, т. е. оценить природу взаимозависимости между наблюдаемыми переменными.

Дополнительная задача корреляционного анализа (являющаяся основной в регрессионном анализе) состоит в оценке уравнений регрессии, где в качестве результативного признака выступает признак, являющийся следствием других признаков (факторов) — причин. Причинно-следственная связь устанавливается из внестатистических соображений,

например из аргументов, касающихся физической природы явлений.

Иногда имеет смысл оценить уравнение регрессии для измерения результативного признака по факторным моделям, несмотря на то, что причинно-следственной связи на самом деле между ними не существует. Здесь причиной могут быть другие факторы, не рассматриваемые в модели, но действующие как на функцию, так и на аргументы уравнения регрессии. Так следует поступать в том случае, когда непосредственное измерение результативного признака затруднительно, но существует тесная корреляционная связь (коэффициент множественной корреляции достаточно близок к единице) между результативным признаком и факторными, измерять и наблюдать которые легче в последующих исследованиях.

Назовем параметр связи в генеральной совокупности значимо отличающимся от нуля (значимым), если гипотеза о равенстве нулю этого параметра отвергается с заданным уровнем значимости α. Если же эта гипотеза принимается, генеральный параметр связи называется незначимым. В корреляционной модели соответствующая связь между величинами считается недоказанной или отсутствующей.

3.2.1. Точечные оценки параметров

Рассмотрим генеральную совокупность с двумя признаками x и y, совместное распределение которых задано плотностью двумерного нормального закона:

$$p = (x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\{Q_2(x, y)\}, \qquad (121)$$

где

$$Q_2(x,y)=\frac{1}{2\sqrt{1-\rho^2}}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right],$$

определяемого пятью параметрами

$$M_x=\mu_x,\ D_x=\sigma_x^2,\ M_y=\mu_y,\ D_y=\sigma_y^2,\ M\left[\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}\right]=\rho,$$

$$\rho^2\neq 1.$$

Имея эти параметры, можно получить уравнения линий регрессии, показывающих изменение условных математических ожиданий в зависимости от изменения соответствующих значений случайных аргументов:

$M_y/x-M_y=\beta_{yx}(x-M_x)$ — прямая регрессии *y* на *x*;

$M_y/x-M_x=\beta_{xy}(y-M_y)$ — прямая регрессии *x* на y;

$\beta_{yx}=\rho\frac{\sigma_y}{\sigma_x}$ – коэффициент регрессии *y* на *x*;

$\beta_{xy}=\rho\frac{\sigma_x}{\sigma_y}$ – коэффициент регрессии *x* на *y*.

Полезно вспомнить, что квадрат коэффициента корреляции ρ^2, т. е. *коэффициент детерминации, в рассматриваемой модели указывает долю дисперсии одной случайной величины, обусловленную вариацией другой. Коэффициент регрессии* β_{yx} *показывает, на сколько единиц своего измерения увеличится* ($\beta>0$) *или уменьшится* ($\beta<0$) *в среднем y(My/x), если x увеличить на единицу своего измерения.*

Задача двумерного корреляционного анализа состоит, прежде всего, в оценке пяти параметров, определяющих генеральную совокупность.

В качестве точечных оценок неизвестных начальных моментов первого и второго порядка генеральной совокупности берутся соответствующие выборочные моменты.

Точечные же оценки неизвестных других параметров получают с помощью формул, аналогичных формулам вычисления самих параметров через генеральные начальные моменты. Таким образом, будем иметь:

$\bar{x}$ — оценка для μ_x,

$\bar{y}$ — оценка для μ_y,

$\bar{x}^2$ — оценка для $M(x^2)$,

$\bar{y}^2$ — оценка для $M(y^2)$,

$\overline{xy}$ — оценка для $M(xy)$.

Откуда

$s_x^2 = \bar{x}^2 - (\bar{x})^2$ — оценка для σ_x^2,

$s_y^2 = \bar{y}^2 - (\bar{y})^2$ — оценка для σ_y^2,

$r = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x s_y}$ — оценка для ρ.

Оценки генеральных коэффициентов регрессии β_{yx} и β_{xy} получаются соответственно по формулам:

$$b_{yx} = r\frac{s_y}{s_x}, \quad b_{xy} = r\frac{s_x}{s_y},$$

откуда оценки уравнений регрессии имеют вид:

$$\overline{y/x} - \bar{y} = b_{yx}(x - \bar{x}), \quad \overline{x/y} - \bar{x} = b_{xy}(y - \bar{y}).$$

При этом $\overline{y/x}$ и $\overline{x/y}$ — обозначения оценок для условных математических ожиданий *My/x* и *Mx/y* генеральной совокупности.

Следует отметить, что вышеприведенные точечные оценки являются состоятельными, а $\overline{x}$ и $\overline{y}$ несмещенными и эффективными. Кроме того, распределение выборочных средних $(\overline{x},\overline{y})$ не зависит от распределения (s_x^2, s_y^2, r). Наконец, выборочный коэффициент корреляции r по абсолютной величине не превосходит единицы.

2.3.3. Приемы вычисления выборочных характеристик

Если объем выборки невелик, то наблюдаемые точки располагают в таблице в порядке их регистрации и обрабатывают по следующей схеме:

x	y	x^2	y^2	xy
.	.	.	.	.
.	.	.	.	.
x_j	y_j	x_j^2	y_j^2	$x_j y_j$
.	.	.	.	.
.	.	.	.	.
$\sum x_j$	$\sum y_j$	$\sum x_j^2$	$\sum y_j^2$	$\sum x_j y_j$

В схеме последовательно заполняют столбцы таблицы результатами операций, указанных сверху. В последней строке вычисляются соответствующие суммы элементов столбцов. Далее используют формулы:

$$\bar{x} = \frac{\sum x_i}{n}; \bar{y} = \frac{\sum y_j}{n}; s_x^2 = \frac{\sum x_j^2}{n} - (\bar{x})^2; s_y^2 = \frac{\sum y_j^2}{n} - (\bar{y})^2;$$

$$r = \frac{\sum x_j y_j - \left(\sum x_j \sum y_j\right)/n}{\left\{\sum x_j^2 - \left(\sum x_j\right)^2 / n \left[\sum y_j^2 - \left(\sum y_j\right)^2 / n\right]\right\}^{1/2}};$$

$$b_{yx} = \frac{\sum x_j y_j - \left(\sum x_j \sum y_j\right)/n}{\sum x_j^2 - \left(\sum x_j\right)^2 / n}; b_{xy} = \frac{\sum x_j y_j - \left(\sum x_j \sum y_j\right)/n}{\sum y_j^2 - \left(\sum y_j\right)^2 / n}.$$

Если выборка многочисленна, то данные группируются путем построения двумерного интервального ряда, корреляционная таблица для которого имеет вид:

x y	…	$(a_k\text{-}b_k]$	…	m_y
. . . $(c_l\text{-}d_l]$ . . . m_x	. . . … . . . …	. . . m_{kl} m_{k*}	. . . … . . . …	. . . m_{*l} . . . n

В таблице m_{kl} — частота прямоугольника, в основании которого лежит полуинтервал $(a_k - b_k]$, а по высоте — $(c_l - d_l]$, т.е. число точек выборки, попавших внутри или на часть границы прямоугольника, задаваемой полуинтервалами. При этом длины интервалов по x одинаковы и равны h_x; то же самое относится и к y (одинаковой длины h_y).

Для вычисления характеристик интервального вариационного ряда переходим к условному дискретному вариационному ряду с условными вариантами

$$x_k' = \frac{x_k - x_0}{h_x}; \; y_l' = \frac{y_l - y_0}{h_y},$$

где x_0, y_0 — рабочие средние — выбираются обычно равными центрам интервалов, лежащих в середине соответствующих одномерных рядов;

x_k — центры интервалов.

Таким образом, условные варианты — целые числа, наименее уклоняющиеся от нуля по абсолютной величине.

Вычисления удобно производить по схеме, последовательно заполняя строки, лежащие ниже таблицы двумерного ряда условных вариантов *(1 ÷ 4)*, и столбцы, лежащие справа от этой таблицы *(1 ÷ 2)*:

x' y'	… x_k …	m_y	$y' m_y^1$	$(y')^2 m_y^2$
. y_l .	. . . … m_{kl} … . . .	. m_{xl} .	. $y_l' m_{*l}$.	. $(y_l')^2 m_{*l}$.
m_x	… m_{k*} …	n	$\sum y_l' m_{*l}$	$\sum (y_l')^2 m_{*l}$
1. $x' m_x$ 2. $(x')^2 m_x$ 3. $\sum y' m_{xy}$ 4. $x' \sum y' m_{xy}$	… $x_k' m_{k*}$ … … $(x_k')^2 m_{k*}$ … … $\sum y_l' m_{kl}$ … … $x' \sum y_l' m_{kl}$ …	$\sum x_k' m_{k*}$ $\sum (x_k')^2 m_{k*}$ $\sum y_l' m_{*l}$ $\sum\sum x_k' y_l' m_{kl}$		

Заметим, что для контроля вычислений можно использовать равенство $\sum\sum x_k' y_l' m_{kl} = \sum y_l' m_{*l}$, т.е. равенство чисел в конце строки 3 и столбца 1.

Далее используются формулы:

$$\bar{x}_{гр} \frac{\sum x_k' m_{k*}}{n} h_x + x_0;\ \bar{y}_{гр} = \frac{\sum \bar{y}_l m_{*l}}{n} h_y + y_0;$$

$$S^2_{x_{гр}} = \left[\frac{\sum\left(x_k^{'}\right)^2 m_{k*}}{n} - \left(\frac{\sum x_k^{'} m_{k*}}{n}\right)^2\right] h_x^2;$$

$$S^2_{y_{гр}} = \left[\frac{\sum\left(y_l^{'}\right)^2 m_{l*}}{n} - \left(\frac{\sum y_l^{'} m_{l*}}{n}\right)^2\right] h_y^2;$$

$$r_{гр} = \frac{n\sum\sum x_k^{'} y_l^{'} m_{kl} - \sum x_k^{'} m_{k*} \sum y_l^{'} m_{*l}}{\left\{\left[n\sum\left(x_k^{'}\right)^2 m_{k*} - \left(\sum x_k m_{k*}\right)^2\right]\left[n\sum\left(y_l^{'}\right)^2 m_{*l} - \left(\sum y_l^{'} m_{*l}\right)^2\right]\right\}^{1/2}};$$

$$byx_{гр} = \frac{n\sum\sum x_k^{'} y_l^{'} m_{kl} - \sum x_k^{'} m_{k*} \sum y_l^{'} m_{*l}}{n\sum\left(y_l^{'}\right) m_{*l} - \left(\sum y_l^{'} m_{*l}\right)^2} \cdot \frac{h_x}{h_y};$$

$$bxy_{гр} = \frac{n\sum\sum x_k^{'} y_l^{'} m_{kl} - \sum x_k^{'} m_{k*} \sum y_l^{'} m_{xl}}{n\sum\left(y_l^{'}\right) m_{*l} - \left(\sum y_l^{'} m_{*l}\right)^2} \cdot \frac{h_x}{h_y}.$$

При группировке вычисленные характеристики могут сильно отличаться от выборочных. Оценки по группированным данным центральных моментов второго порядка s_x^2 и s_y^2 можно **улучшить поправками Шеппарда:**

$$s_x^2 \cong s^2_{x_{гр}} - \frac{1}{12} h_x^2;\ s_y^2 \cong s^2_{y_{гр}} - \frac{1}{12} h_y^2. \qquad (122)$$

Эти поправки часто сглаживают ошибки, возникающие от группировки, если длина интервала *(h)* не превосходит восьмой части размаха соответствующего признака.

2.3.4. Проверка значимости параметров связи

В двумерной модели параметрами связи являются коэффициент корреляции ρ (или квадрат, называемый коэффициентом детерминации) и коэффициенты регрессии

β_{yx} *и* β_{xy}.

Заметим, что в двумерной модели достаточно проверить значимость только коэффициента корреляции. Если коэффициент корреляции незначим, то признаки *x* и *y* считаются независимыми в генеральной совокупности.

Статистика *r*, вычисляемая для выборки из двумерной нормально распределенной совокупности с $\rho = 0$, связана со статистикой *t*, имеющей распределения Стьюдента с $\nu=n-2$ степенями свободы, формулой

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}. \qquad (123)$$

Зная границы для t, соответствующие обычным уровням значимости (α=10%, 5%, 2%, 1%), можно получить границы для *r*, воспользовавшись этой формулой. Границы для *r* табулированы. Таким образом, для проверки гипотезы $H_{0:}\ \rho=0$ по данным α и $\nu = n-2$ находим $r_{табл}$. *Если* $|r_{набл}| > r_{табл}$, то гипотеза *Но* отвергается с вероятностью ошибки α, если же $|r_{набл}| < r_{табл,}$ то гипотеза не отвергается. При $\nu > 100$ для проверки $H_{0:}\ \rho=0$ следует пользоваться нормированным нормальным законом распределения статистики

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{\nu}.$$

$$r\sqrt{n-1}.$$

Если наблюдаемая величина (*t* или $r\sqrt{n-1}$) расположена в доверительном интервале $[-t_{1-\alpha},\ t_{1-\alpha}]$, то гипотеза H_0 не отвергается; в противном случае H_0 отвергается с уровнем значимости α.

2.3.5. Интервальные оценки параметров связи

Для значимых параметров связи имеет смысл найти интервальные оценки.
При нахождении доверительного интервала для коэффициента корреляции ρ используют статистику, введенную Фишером:

$$z_r = \frac{1}{2}\ln\frac{r+1}{1-r},$$

которая при $n>10$ распределена приблизительно нормально с генеральным средним $MZ_r \approx \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$ и дисперсией $Dz_r \approx \frac{1}{n-3}$.

Тогда доверительный интервал, оценивающий MZ_r с надежностью $\gamma=1-\alpha$, имеет вид

$$Z_r - t_\gamma\sqrt{\frac{1}{n-3}} \le MZ_r \le Z_r + t_\gamma\sqrt{\frac{1}{n-3}}, \qquad (124)$$

где t_γ — находится по таблицам интеграла Лапласа для данного γ (или $\gamma=1-\alpha$).
Для перехода от Z к ρ имеется таблица, составленная Фишером и Иейтсом, после использования которой получаем интервальную оценку с надежностью γ вида

$$r_{\min} \le \rho \le r_{\max},$$

где r_{min} и r_{max} выбираются с учетом того, что Z_r — функция нечетная. При этом поправочным членом $\frac{\rho}{2(n-1)}$ у MZ_r пренебрегают.

Если коэффициент корреляции значим, то коэффициенты регрессии также значимо отличаются от нуля (с тем же уровнем α). Интервальные оценки для них получаются по формулам:

$$|t| \le t(\gamma,\nu);\ t = \left(b_{yx} - \beta_{yx}\right)\frac{s_x\sqrt{n-2}}{s_y\sqrt{1-r^2}};\ t = \left(b_{xy} - \beta_{xy}\right)\frac{s_y\sqrt{n-2}}{s_x\sqrt{1-r^2}},$$

где t имеет распределение Стьюдента с $\nu = n-2$ степенями свободы.

Переход от неравенства $|t| \le t(\gamma,\nu)$ к интервальным оценкам для коэффициента регрессии осуществляется с помощью тождественных алгебраических преобразований.

Для значимого коэффициента корреляции ρ некоторые авторы рекомендуют более предпочтительную оценку, чем r:

$$r\left(1+\frac{1-r^2}{2(n-4)}\right).$$

Предпочтительной оценкой ρ^2 является выражение

$$\frac{(n-1)r^2-1}{n-2}.$$

Этими точечными оценками следует пользоваться при небольших объемах n выборки.

Кроме нахождения интервальной оценки для ρ, с помощью преобразования

$$Z_r = \frac{1}{2}\ln\frac{1+r}{1-r}$$

можно решить следующие задачи.

1. Проверить, согласуется ли выборочный коэффициент корреляции r с предполагаемым значением генерального коэффициента корреляции ρ_0. Для этого, взяв уровень значимости α, проверяем, попадает ли абсолютная величина разности $|Z_r - Z_\rho|$ в интервал $[0, t1-\alpha/\sqrt{n-3}]$. Если попадает, то гипотеза $H_0: \rho = \rho_0$ не отвергается. В противном случае отвергается с уровнем α.

1. Проверить гипотезу об однородности коэффициентов корреляции. Пусть $r_1, r_2, ..., r_k$. — коэффициенты корреляции, полученные из k нормально распределенных совокупностей по выборкам с объемами $n_1, n_2, ..., n_k$. Проверяется гипотеза

$$H_0 : \rho_1 = \rho_2 = ... = \rho_k = \rho.$$

Статистика

$$\sum_{i=1}^{k} \frac{\left(z_{r_i} - z_\rho\right)^2}{1/(n_i - 3)}$$

имеет тогда распределение χ^2 с k степенями свободы. Если заменить z_ρ на среднее арифметическое

$$\bar{z}_r = \frac{\sum z_i n_i}{\sum n_i},$$

то получим, что

$$\sum_{i=1}^{k} \frac{\left(z_{r_i} - \bar{z}_r\right)^2}{1/(n_i - 3)}$$

распределена по закону χ^2 с $\nu=k-1$ степенями свободы. Если теперь для заданных a и $\nu=k-1$

$$x^2_{набл} < \sum_{i=1}^{k} \frac{\left(z_{r_i} - \bar{z}_r\right)^2}{1/(n_i - 3)},$$

то гипотеза однородности отвергается с уровнем α. В противном случае гипотеза H_0 не отвергается.

В случае принятия гипотезы однородности предпочтительной точечной оценкой ρ является значение r, полученное обратным преобразованием из z_r.

2.4. Трехмерная модель

2.4.1. Основные параметры модели

Для изучения основных задач и особенностей корреляционного анализа удобно рассматривать генеральную совокупность трех признаков *x*, *y* и *z*.

Трехмерная непрерывная случайная величина *(x, y, z)* называется **нормально распределенной,** если плотность совместного распределения одномерных случайных величин *x*, *y* и *z* задается в виде

$$\rho(x,y,z)=(2\pi)^{-\frac{3}{2}}\left[\sigma_x^2\sigma_y^2\sigma_z^2\left|\Re_3\right|\right]^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}u^T\Re_3^{-1}u\right\},$$

где $\Re_3=\begin{pmatrix}1 & \rho_{xy} & \rho_{xz}\\ \rho_{yz} & 1 & \rho_{yz}\\ \rho_{zx} & \rho_{zy} & 1\end{pmatrix}$ — симметрическая положительно определенная матрица парных коэффициентов корреляции, соответствующих частным двумерным распределениям случайных величин *(x,y)*, *(x, z)* и *(y,z)*;

$\left|\Re_3\right|$ — определитель матрицы $\Re_3$, обобщенная дисперсия случайной величины *(x, y, z);*

$$\left|\Re_3\right|=1+2\rho_{xy}\rho_{xz}\rho_{yz}-\rho_{xz}^2-\rho_{yz}^2-\rho_{xy}^2>0;$$

$\Re_3^{-1}=\begin{pmatrix}\rho^{(11)} & \rho^{(12)} & \rho^{(13)}\\ \rho^{(21)} & \rho^{(22)} & \rho^{(23)}\\ \rho^{(31)} & \rho^{(32)} & \rho^{(33)}\end{pmatrix}$ - матрица, обратная $\Re_3$:

$\rho^{(ij)} = \frac{\Re_{ji}}{|\Re_3|};\ \Re_{ji} = (-1)^{j+1} M_{ji},\ M_{ji}$ - минор матрицы $\Re_3$,

дополнительный к элементу ρ_{ji},

j,i=1,2,3;

$$\rho^{(11)} = \frac{1}{|\Re_3|}(-1)^{1+1}\begin{pmatrix} 1 & \rho_{yz} \\ \rho_{yx} & 1 \end{pmatrix} = \frac{1-\rho_{yz}^2}{|\Re_3|};$$

$$\rho^{(12)} = \frac{1}{|\Re_3|}(-1)^{1+2}\begin{vmatrix} \rho_{xy} & \rho_{xz} \\ \rho_{yz} & 1 \end{vmatrix} = \frac{-\rho_{xy} + \rho_{xz}\rho_{yz}}{|\Re_3|};$$

$$\rho^{(13)} = \frac{1}{|\Re_3|}(-1)^{1+3}\begin{vmatrix} \rho_{xy} & \rho_{xz} \\ 1 & \rho_{yz} \end{vmatrix} = \frac{\rho_{xy}\rho_{yz} - \rho_{xz}}{|\Re_3|};$$

$$\rho^{(22)} = \frac{1}{|\Re_3|}(-1)^{2+2}\begin{vmatrix} 1 & \rho_{xz} \\ \rho_{yz} & 1 \end{vmatrix} = \frac{1-\rho_{xz}^2}{|\Re_3|};$$

$$\rho^{(23)} = \frac{1}{|\Re_3|}(-1)^{2+3}\begin{vmatrix} 1 & \rho_{xz} \\ \rho_{yz} & \rho_{yz} \end{vmatrix} = \frac{-\rho_{yz} + \rho_{xz}\rho_{xy}}{|\Re_3|};$$

$$\rho^{(33)} = \frac{1}{|\Re_3|}(-1)^{3+3}\begin{vmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{vmatrix} = \frac{1-\rho_{xy}^2}{|\Re_3|},$$

матрица $\Re_3^{-1}$— симметрическая, положительно определенная;

$$u = \begin{pmatrix} \dfrac{x-\mu_x}{\sigma_x} \\ \dfrac{y-\mu_y}{\sigma_y} \\ \dfrac{z-\mu_z}{\sigma_z} \end{pmatrix}$$ - вектор значений нормированных случайных величин x, y, z;

u^T — транспонированный вектор *u:*

$$u^T \Re_3^{-1} u = \begin{pmatrix} \dfrac{x-\mu_x}{\sigma_x} & \dfrac{y-\mu_y}{\sigma_y} & \dfrac{z-\mu_z}{\sigma_z} \end{pmatrix} \begin{pmatrix} \rho^{(11)} & \rho^{(12)} & \rho^{(13)} \\ \rho^{(21)} & \rho^{(22)} & \rho^{(23)} \\ \rho^{(31)} & \rho^{(32)} & \rho^{(33)} \end{pmatrix} \times$$

$$\times \begin{pmatrix} \dfrac{x-\mu_x}{\sigma_x} \\ \dfrac{y-\mu_y}{\sigma_y} \\ \dfrac{z-\mu_z}{\sigma_z} \end{pmatrix} = \left[\frac{x-\mu_x}{\sigma_x} \rho^{(11)} + \frac{y+\mu_y}{\sigma_y} \rho^{(12)} + \frac{z-\mu_z}{\sigma_z} \rho^{(13)} \right] \times$$

$$\times \frac{x-\mu_x}{\sigma_x} + + \left[\frac{x-\mu_x}{\sigma_x} \rho^{(12)} + \frac{y-\mu_y}{\sigma_y} \rho^{(22)} + \frac{z-\mu_z}{\sigma_z} \rho^{(23)} \right] \times$$

$$\times \frac{y-\mu_y}{\sigma_y} + \left[\frac{x-\mu_x}{\sigma_x} \rho^{(13)} + \frac{y-\mu_y}{\sigma_y} \rho^{(32)} + \frac{z-\mu_z}{\sigma_z} \rho^{(33)} \right] \times$$

$$\times \frac{z-\mu_z}{\sigma_z}.$$

Таким образом, трехмерная нормально распределенная случайная величина определяется *девятью параметрами:* тремя математическими ожиданиями:

$$M_x = \mu_x,\ M_y = \mu_y,\ M_z = \mu_z;$$

тремя дисперсиями (или средними квадратическими отклонениями):

$$D_x = \sigma_x^2, D_y = \sigma_y^2, D_z = \sigma_z^2\ \ (\sigma_x, \sigma_y, \sigma_z);$$

тремя парными коэффициентами корреляции:

$$\rho_{xy} = M\left[\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}\right],$$

$$\rho_{xz} = M\left[\frac{x-\mu_x}{\sigma_x}\cdot\frac{z-\mu_z}{\sigma_z}\right],$$

$$\rho_{yz} = M\left[\frac{y-\mu_y}{\sigma_y}\cdot\frac{z-\mu_z}{\sigma_z}\right].$$

Следует отметить, что частные одномерные *(x, y* и *z),* двумерные *((x,y),(x,z)* и *(y,z))* распределения компонент, а также условные распределения при фиксировании одной *((x,y)/z, (x,z)/y, (y,z)/x)* и двух *(x/y,z; y/xz; z/x,y)* компонент являются нормальными. Поэтому поверхности и линии регрессии являются плоскостями и прямыми соответственно.

Для трехмерной (и других многомерных) корреляционной модели важную роль играют частные и множественные коэффициенты корреляции или детерминации (коэффициент детерминации равен квадрату соответствующего коэффициента корреляции).

Частным коэффициентом корреляции между *x* и *y* при фиксированных остальных компонентах (т. е. *z*) является выражение

$$\rho_{xy/z} = -\frac{\Re_{12}}{\left(\Re_{11}\cdot\Re_{12}\right)^{1/2}} = \frac{\rho_{xy}-\rho_{xz}\rho_{yz}}{\sqrt{\left(1-\rho_{xz}^2\right)\left(1-\rho_{yz}^2\right)}}. \quad (125)$$

Остальные частные коэффициенты корреляции $\rho_{xz/y}$ и $\rho_{yz/x}$ определяют путем замены соответствующих индексов в приведенных формулах.

Для нормального распределения частный коэффициент корреляции $\rho_{xy/z}$ совпадает с парным коэффициентом корреляции между величинами *x* и *y* при фиксированном *z* (в двумерном условном распределении *((x,y)/z). Частный коэффициент корреляции обладает всеми свойствами парного коэффициента корреляции.*

Он служит показателем линейной связи между двумя переменными случайными величинами независимо от влияния остальных случайных переменных. Если частный коэффициент детерминации меньше, чем соответствующий парный коэффициент детерминации, то взаимозависимость между двумя величинами обусловлена частично (или целиком при равенстве нулю частного коэффициента детерминации) воздействием на эту пару остальных, фиксируемых, случайных величин. Если же, наоборот, частный коэффициент детерминации больше соответствующего парного, то фиксируемые величины ослабляют, затушевывают связь.

Множественный коэффициент корреляции между одной величиной *z* и двумя другими величинами *(x,y)* определяется по формуле

$$\rho_z = \rho_{z/xy} = \sqrt{1 - \frac{|\Re_3|}{\Re_{11}}} = \sqrt{\frac{\rho_{zx}^2 + \rho_{zy}^2 - 2\rho_{xy}\rho_{zx}\rho_{zy}}{1 - \rho_{xy}^2}}.$$

(126)

Для трехмерной нормально распределенной случайной величины *(x,y,z)* множественный коэффициент корреляции является мерой связи между одной случайной величиной и двумя остальными. Он заключен между нулем и единицей. При $\rho_z = 1$ связь между величинами *z* и *(x,y)* является функциональной, линейной: точки *(x,y,z)* расположены в

плоскости регрессии *z* на *(x,y)*. При $\rho_z = 0$ одномерная случайная величина *z* и двумерная случайная величина *(x,y)* являются независимыми (в силу нормальности распределения). *Множественный коэффициент детерминации* ρ_z^2 *показывает долю дисперсии случайной величины z, обусловленную изменением случайных величин (x,y)*.

Из определяющей ρ_z формулы можно получить следующие неравенства:

$$\rho_z \geq |\rho_{zx}|,\ \rho_z \geq |\rho_{zy}|,\ \rho_z \geq |\rho_{zx/y}|,\ \rho_z \geq |\rho_{zy/x}|.$$

Отсюда можно заметить, что коэффициент множественной корреляции может только увеличиться, если в модель включать дополнительные признаки — случайные величины, и не увеличиться, если из имеющихся признаков производить исключение.

Далее, если $\rho_z = 0$, то $\rho_{zx} = \rho_{zy} = \rho_{zx/y} = \rho_{zy/x} = 0$.

Если, например, $\rho_{xy}^2 \leq \rho_{xz}^2$ и $\rho_{xy}^2 \leq \rho_{yz}^2$, то

$\rho_z^2 \geq \rho_x^2, \rho_z^2 \geq \rho_y^2$ и $\rho_{xy/z}^2 \leq \rho_{xz/y}^2$, $\rho_{xy/z}^2 \leq \rho_{yz/x}^2$.

Последние неравенства можно получить исходя из формул:

$$\rho_{xy/z}^2 = \begin{cases} 1 - \dfrac{1-\rho_x^2}{1-\rho_{xz}^2} \\ 1 - \dfrac{1-\rho_y^2}{1-\rho_{yz}^2} \end{cases}; \rho_{xz/y}^2 = \begin{cases} 1 - \dfrac{1-\rho_y^2}{1-\rho_{xy}^2} \\ 1 - \dfrac{1-\rho_z^2}{1-\rho_{yz}^2} \end{cases}; \rho_{yz/x}^2 = \begin{cases} 1 - \dfrac{1-\rho_y^2}{1-\rho_{xy}^2} \\ 1 - \dfrac{1-\rho_z^2}{1-\rho_{xz}^2} \end{cases}.$$

Таким образом, наибольшему множественному коэффициенту детерминации соответствуют большие частные коэффициенты детерминации (например, ρ_x^2 соответствуют $\rho_{xz/y}^2$ и $\rho_{xy/z}^2$).

Приведем некоторые характеристики, подлежащие корреляционному анализу трехмерной случайной величины. При этом будем рассматривать лишь по одному условному распределению (двумерному и одномерному), так как остальные совпадают с рассматриваемыми с точностью до перестановки букв.

Условное распределение при заданном z

Так как это двумерное нормальное распределение *(x,y)/z*, то оно определяется пятью параметрами (двумя условными математическими ожиданиями $\mu_{x/z}$, и $\mu_{y/z}$, двумя условными дисперсиями $\sigma^2_{x/z}$ и $\sigma^2_{y/z}$; условным коэффициентом корреляции $\rho_{xy/z}$):

$$\mu_{x/z} = \mu_x + \rho_{zx}\frac{\sigma_x}{\sigma_z}(z-\mu_z);\ \mu_{y/z} = \mu_y + \rho_{zy}\frac{\sigma_y}{\sigma_z}(z-\mu_z);$$

$$\sigma^2_{x/z} = \sigma^2_x\left(1-\rho^2_{zx}\right);\ \sigma^2_{y/z} = \sigma^2_y\left(1-\rho^2_{zx}\right);\ \rho_{yx/z} = \frac{\rho_{xy}-\rho_{xz}\rho_{yz}}{\left[\left(1-\rho^2_{xz}\right)\left(1-\rho^2_{yz}\right)\right]^{1/2}}.$$

Форма зависимости выражается следующими линиями регрессии в плоскости $Z=\overline{z}$:

$$M(y/x)/z - \mu_{y/z} = \beta_{yx/z}(x-\mu_{x/z});$$
$$M(x/y)/z - \mu_{x/z} = \beta_{xy/z}(y-\mu_{y/z}).$$

Коэффициенты частной регрессии имеют вид:

$$\beta_{yx/z} = \rho_{xy/z}\frac{\sigma_{y/z}}{\sigma_{x/z}} = \frac{\beta_{yx}-\beta_{yz}\beta_{zx}}{1-\beta_{xz}\beta_{zx}};$$

$$\beta_{xy/z} = \rho_{xy/z}\frac{\sigma_{x/z}}{\sigma_{y/z}} = \frac{\beta_{xy}-\beta_{xz}\beta_{zy}}{1-\beta_{yz}\beta_{zy}};$$

(127)

причем

$$\rho^2_{xy/z} = \beta_{xy/z}\beta_{yx/z}.$$

Условные средние квадратические отклонения (при двух условиях), характеризующие рассеяние относительно указанных линий регрессии и совпадающие с остаточными средними квадратическими отклонениями, определяются формулами:

$$\sigma_{y/zx} = \sigma_{y/z}\sqrt{1-\rho^2_{xy/z}} = \sigma_{y/x}\sqrt{1-\rho^2_{yz/x}};$$

$$\sigma_{x/yz} = \sigma_{x/z}\sqrt{1-\rho^2_{xy/z}} = \sigma_{x/y}\sqrt{1-\rho^2_{xz/y}}.$$

Центр условного двумерного распределения *(M(x/y)/z, M{y/x)/z)* при изменении *Z* описывает прямую в пространстве 0_{xyz}, в то время, как условные дисперсии $\sigma^2_{x/z}, \sigma^2_{y/z}$ и условный коэффициент корреляции $\rho_{xy/z}$ остаются постоянными.

Условное распределение при заданном (x, y)

Это распределение *z/(x,y)* является одномерным и определяется своими математическим ожиданием и дисперсией (естественно условными):

$$M_z/(x,y) = M(z/x)/y = M(z/y)/x; \quad D_z(x,y) = \sigma^2_{z/xy}.$$

Если точку *(x,y)* менять, то будем иметь плоскость регрессии *z* на *(x,y)*

$$M_z(x,y) - \mu_z = \beta_{zx/y}(x-\mu_x) + \beta_{zy/x}(y-\mu_y)$$

и остаточную дисперсию относительно плоскости регрессии (совпадающую с условной дисперсией)

$$\sigma^2_{z/xy} = \sigma^2_{z/y}\left(1-\rho^2_{zx/y}\right) = \sigma^2_{z/x}\left(1-\rho^2_{yz/x}\right).$$

Коэффициент множественной регрессии (совпадающий с соответствующим коэффициентом частной регрессии), например, $\beta_{zx/y}$, показывает, на сколько единиц своего

измерения изменится признак *z* в среднем, если признак *x* изменится на единицу своего измерения, а остальные признаки не изменятся. Таким образом, коэффициент регрессии может выступать в качестве норматива.

Множественный коэффициент корреляции ρ_z можно вычислить в силу линейности регрессии и как корреляционное отношение *z* на *(xy)*:

$$\rho_z = \eta_z = \sqrt{1 - \frac{D_{остZ}}{D_{общZ}}} = \sqrt{1 - \frac{\sigma^2_{z/xy}}{\sigma^2_z}}.$$

Если, например, $\rho_{zx/y} = 0$, то из последней формулы следует:

$$\rho^2_z = 1 - \frac{\sigma^2_{z/xy}}{\sigma^2_z} = 1 - \frac{\sigma^2_{z/y}}{\sigma^2_z} = \rho^2_{xy}.$$

2.4.2. Оценивание и проверка значимости параметров

Пусть дана выборка объемом из трехмерной нормально распределенной генеральной совокупности с признаками *x, y* и *z*:

$$(x_1, y_1, z_1),\ (x_2, y_2, z_2),\ldots,\ (x_n, y_n, z_n).$$

Обработку данных будем производить, руководствуясь таблицей приведенной ниже:

x	*y*	*z*	x^2	y^2	z^2	*xy*	*xz*	*yx*
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
x_j	y_j	z_j	x_j^2	y_j^2			$x_j z_j$.	$y_j z_j$.

. . . .	. . .	. . .	. . .	. . .	z_j^2 . . .	$x_j y_j$ . . .	. .	. .
$y\sum x$	$\sum y$	$\sum z$	$\sum x^2$	$\sum y^2$	$\sum z^2$	$\sum xy$	$\sum xz$	$\sum yz$

Точечные оценки девяти генеральных параметров μ_x, μ_y, μ_z, $\sigma_x^2, \sigma_y^2, \sigma_z^2$ ρ_{xy}, ρ_{xz} и ρ_{yz} можно вычислить по формулам:

$$\bar{x} = \frac{1}{n}\sum x;\ s_x^2 = \frac{\sum x^2}{n} - (\bar{x})^2;\ r_{xy} = \frac{\frac{1}{n}\sum xy - \bar{x}\cdot\bar{y}}{s_x s_y};$$

$$\bar{y} = \frac{1}{n}\sum y;\ s_y^2 = \frac{\sum y^2}{n} - (\bar{y})^2;\ r_{xz} = \frac{\frac{1}{n}\sum xz - \bar{x}\cdot\bar{z}}{s_x s_z}; \quad (128)$$

$$\bar{z} = \frac{1}{n}\sum z;\ s_z^2 = \frac{\sum z^2}{n} - (\bar{z})^2;\ r_{yz} = \frac{\frac{1}{n}\sum yz - \bar{y}\cdot\bar{z}}{s_y s_z}.$$

Затем вычисляются оценки условных средних квадратических отклонений при фиксировании одной компоненты, частных коэффициентов корреляции, условных средних квадратических отклонений при двух фиксированных компонентах и множественных коэффициентов корреляции, используя формулы, соответствующие формулам для вычисления параметров генеральной совокупности:

$$s_{x/y} = s_x\sqrt{1-r_{xy}^2};\ s_{x/z} = s_x\sqrt{1-r_{xz}^2};\ s_{y/z} = s_y\sqrt{1-r_{yz}^2};$$

$$s_{x/y} = s_x\sqrt{1-r_{xy}^2};\ s_{x/z} = s_x\sqrt{1-r_{xz}^2};\ s_{y/z} = s_y\sqrt{1-r_{yz}^2};$$

$$\kappa_{xy/z} = \frac{r_{xy} - r_{xz} r_{yz}}{\left[\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)\right]^{1/2}};$$

$$\kappa_{xz/y} = \frac{r_{xz} - r_{xy} r_{yz}}{\left[\left(1 - r_{xy}^2\right)\left(1 - r_{yz}^2\right)\right]^{1/2}};$$

(129)

$$\kappa_{yz/x} = \frac{r_{yz} - r_{xy} r_{xz}}{\left[\left(1 - r_{xy}^2\right)\left(1 - r_{xz}^2\right)\right]^{1/2}};$$

$$\kappa_{x/yz} = s_{x/y}\sqrt{1 - r_{xz/y}^2};\ s_{y/xz} = s_{y/z}\sqrt{1 - r_{xy/z}^2};$$

(130)

$$s_{z/xy} = s_{z/x}\sqrt{1 - r_{yz/x}^2};$$

$$r_x = \sqrt{1 - \frac{s_{x/yz}^2}{s_x^2}};\ r_y = \sqrt{1 - \frac{s_{y/xz}^2}{s_y^2}};\ r_z = \sqrt{1 - \frac{s_{z/xy}^2}{s_z^2}}.$$

Проверка значимости множественного коэффициента детерминации ρ^2_M (следовательно, и ρ_M) осуществляется с помощью *F*-распределения. Вычисляется

$$F_{набл} = \frac{r_м^2 / 2}{(1 - r_м^2)/(n-3)}. \qquad (131)$$

Затем с заданным уровнем значимости α и числами степеней свободы $vi=2$ (числителя) и $v_1=n-3$ (знаменателя) находят $F_{табл}$. Если $F_{набл} > F_{табл}$, то гипотеза $H0: \rho^2_M = 0$ отвергается с вероятностью ошибки α, т. е. $\rho^2_{\text{М}}$ значимо отличается от нуля. Если коэффициент незначим, связь между случайной величиной *Z* и случайной величиной *(x,y)* отсутствует.

Конечно, проверку значимости коэффициентов связи начинать с частных коэффициентов корреляции не

обязательно. Можно в некоторых случаях сократить такую проверку, например, если ρ_z незначим, то коэффициенты $\rho_{zx/y}$ и $\rho_{zy/x}$ становятся незначимыми. Далее, если $\rho_{zx/y}$ незначим, то $\rho_z=|\rho_{z/y}|$ (множественный коэффициент корреляции незначимо отличается от абсолютной величины парного коэффициента корреляции).

Для значимых множественных коэффициентов корреляции можно получить оценки уравнения регрессии.

Например, пусть ρ_z значимо отличается от нуля, тогда оценкой соответствующего уравнения регрессии служит

$$\overline{z/(x,y)} - \bar{z} = b_{zx/y}(x-\bar{x}) + b_{yz/x}(y-\bar{y}). \quad (132)$$

При этом коэффициенты регрессии вычисляются по формулам:

$$b_{zx/y} = r_{zx/y}\frac{s_{z/y}}{s_{x/y}};\ b_{zy/x} = r_{zy/x}\frac{s_{z/x}}{s_{y/x}} \quad (133)$$

и $\overline{z/(x,y)}$ является оценкой *Mz/(x,y)*.

Напомним, что если какой-либо частный коэффициент корреляции незначим, то соответствующий коэффициент плоскости регрессии также незначим. Поэтому, если позволяют условия практического анализа, с точки зрения надежности статистических выводов, предпочтительнее рассматривать модель взаимозависимости признаков такую, для которой множественный коэффициент детерминации — наибольший (и, конечно, значимый): ему соответствует максимальное число значимых частных коэффициентов детерминации (корреляции).

Для значимых параметров связи представляет интерес найти интервальную оценку с надежностью $\gamma = 1 - \alpha$.

Интервальная оценка для $\rho_{частн}$ находится с помощью статистики Фишера:

$$Z_r = Z(r) = \frac{1}{2}\ln\frac{1+r}{1-r}.$$

По таблице указанного преобразования находят величину $Z/\rho_{частн}$. Затем вычисляют точность интервальной оценки для *MZ,* воспользовавшись тем фактом, что статистика *Z(r)* распределена приближенно нормально с параметрами $MZ \cong \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ и $DZ \cong \frac{1}{n-4}$:

$$\Delta Z = t_\gamma \frac{1}{\sqrt{n-4}},$$

где t_γ является решением уравнения $\Phi(t) = \gamma$ и находится по таблице интегральной функции Лапласа.

Затем вычисляются границы интервальной оценки для *MZ* по формуле

$$Z(r_{част}) \pm \Delta Z,$$

и, наконец, доверительные границы для $\rho_{част}$ получают по таблице обратного преобразования Фишера.

Для значимого множественного коэффициента корреляции интервальная оценка также находится с помощью Z-преобразования Фишера, с дисперсией, приблизительно равной $\frac{1}{n}$ для достаточно больших значений *n.*

Имеются графики и таблицы (Эзекиела и Фокса; К. Крамера) для получения интервальных оценок ρ_M^2 по значениям r_M^2.

Определение доверительных интервалов для коэффициентов плоскости регрессии производится исходя из статистик:

$$t = \frac{(b_{zx/y} - \beta_{zx/y}) s_{x/y}\sqrt{n-3}}{s_{z/y}\sqrt{1-r_{zx/y}^2}};$$

(134)

$$t = \frac{(b_{zy/x} - \beta_{zy/x}) s_{y/x}\sqrt{n-3}}{s_{z/x}\sqrt{1-r_{zy/x}^2}},$$

которые имеют Z-распределение Стьюдента с $v=n-3$ степенями свободы. Для этого достаточно решить относительно оцениваемого коэффициента регрессии неравенство $|t| \le t(\alpha,n-3)$, где $t(\alpha,n-3)$ находится по таблице Стьюдента.

Для значимых частных и множественных коэффициентов детерминации можно указать более предпочтительные точечные оценки, чем выборочные коэффициенты, например:

$\frac{(n-2)r^2_{xy/z}-1}{n-3}$ - оценка для $\rho^2_{xy/z}$;

$\frac{(n-1)r^2_z-2}{n-3}$ - оценка для ρ_z.

3. МЕТОДЫ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ

3.1. Классификация без обучения. Кластерный анализ

3.1.1. Основные понятия

В статистических исследованиях группировка первичных данных является основным приемом решения задачи классификации, а значит и основой всей дальнейшей работы с собранной информацией.

Традиционно эта задача решается следующим образом. Из множества признаков, описывающих объект, отбирается один, наиболее информативный с точки зрения исследователя, и производится группировка в соответствии со значениями данного признака. Если требуется провести классификацию по нескольким признакам, ранжированным между собой по степени важности, то сначала производится классификация по первому признаку, затем каждый из полученных классов разбивается на подклассы по второму признаку и т. д. Подобным образом строится большинство комбинационных статистических группировок.

В тех случаях, когда упорядочить классификационные признаки не представляется возможным, применяется наиболее простой метод многомерной группировки — **создание интегрального показателя (индекса),** функционально зависящего от исходных признаков, с последующей классификацией по этому показателю.

Развитием этого подхода является вариант классификации по нескольким обобщающим показателям (главным компонентам), полученным с помощью **методов факторного анализа.**

При наличии нескольких признаков (исходных или обобщенных) задача классификации может быть решена

методами кластерного анализа, которые от других методов многомерной классификации отличаются отсутствием обучающих выборок, т. е. априорной информации о распределении генеральной совокупности, которая представляет собой вектор *X.*

Различия между схемами решения задач классификации во многом определяются тем, что понимают под понятиями "сходство" и "степень сходства".

После того как сформулирована цель работы, необходимо попытаться определить **критерии качества, целевую функцию,** значения которой позволят сопоставить различные схемы классификации.

В экономических исследованиях целевая функция, как правило, должна минимизировать некоторый **параметр,** определенный на множестве объектов (например целью классификации оборудования может явиться группировка, минимизирующая совокупность затрат времени и средств на ремонтные работы).

В случаях, когда формализовать цель задачи не удается, критерием качества классификации может служить **возможность содержательной интерпретации найденных групп.**

Рассмотрим следующую задачу. Пусть исследуется совокупность n объектов, каждый из которых характеризуется по k замеренным на нем признакам X. Требуется разбить эту совокупность на однородные в некотором смысле группы (классы). При этом практически отсутствует априорная информация о характере распределения измерений X внутри классов.

Полученные в результате разбиения группы обычно называются **кластерами** (от англ. cluster — группа элементов, характеризуемых каким-либо общим свойством), а также таксонами (от англ. taxon — систематизированная группа любой категории) или образами. Методы нахождения кластеров называются **кластер-анализом** (соответственно

численной таксономией или распознаванием образов с самообучением).

При этом с самого начала необходимо четко представить, какая из двух задач классификации подлежит решению. Если решается **обычная задача типизации,** то совокупность наблюдений разбивают на сравнительно небольшое число областей группирования (например интервальный вариационный ряд в случае одномерных наблюдений) так, чтобы элементы одной такой области по возможности находились друг от друга на небольшом расстоянии.

Решение **другой задачи типизации** заключается в определении естественного расслоения исходных наблюдений на четко выраженные кластеры, лежащие друг от друга на некотором расстоянии.

Если первая задача типизации всегда имеет решение, то при второй постановке может оказаться, что множество исходных наблюдений не обнаруживает естественного расслоения на кластеры, т. е. образует один кластер.

Несмотря на то, что многие методы кластерного анализа довольно элементарны, применение методов кластерного анализа стало возможным только в 80-е годы с возникновением и развитием вычислительной техники. Это объясняется тем, что эффективное решение задачи поиска кластеров требует большего числа арифметических и логических операций. Рассмотрим три различных подхода к проблеме кластерного анализа: эвристический, экстремальный и статистический.

Эвристический подход характеризуется отсутствием формальной модели изучаемого явления и критерия для сравнения различных решений. Его основой является алгоритм, построенный исходя из интуитивных соображений.

При **экстремальном подходе** также не формулируется исходная модель, а задается критерий, определяющий качество разбиения на кластеры. Такой подход особенно полезен, если цель исследования четко определена. В этом

случае качество разбиения может измеряться эффективностью выполнения цели.

Основой **статистического подхода** решения задачи кластерного анализа является вероятностная модель исследуемого процесса. Статистический подход особенно удобен для теоретического исследования проблем, связанных с кластерным анализом. Кроме того, он дает возможность ставить задачи, связанные с воспроизводимостью результатов кластерного анализа.

Рассмотрим формы представления исходных данных и определение мер близости.

В задачах кластерного анализа обычной формой представления исходных данных служит **прямоугольная таблица,** каждая строка которой представляет результат измерения k рассматриваемых признаков на одном из обследованных объектов:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}.$$

В конкретных ситуациях может представлять интерес как группировка объектов, так и группировка признаков. В случаях, когда разница между этими двумя задачами несущественна, например при описании некоторых алгоритмов, мы будем пользоваться только термином "объект", подразумевая в этом понятии и "признак".

Числовые значения, входящие в матрицу X, могут соответствовать трем типам переменных: количественным, ранговым и качественным. **Количественные** переменные обладают свойством упорядоченности и над ними можно производить арифметические операции. Значения **ранговых** переменных тоже упорядочены, и их можно пронумеровать натуральными числами. Однако использование этих чисел в

арифметических операциях будет некорректным. **Качественными** называются переменные, принимающие два (дихотомные) или более значений. Этим значениям также можно поставить в соответствие некоторые числа, которые, однако, не будут отражать какой-либо упорядоченности значений качественной переменной. Исключением являются дихотомные переменные, два значения которых (как правило, они обозначаются числами 0 и 1) можно считать упорядоченными.

Желательно, чтобы таблица исходных данных соответствовала одному типу переменных. В противном случае разные типы переменных стараются свести к какому-то одному типу переменных. Например, все переменные можно свести к дихотомным, используя следующую процедуру. Количественные переменные переводят в ранговые, разбивая области значений количественной переменной на интервалы, которые затем нумеруются числами натурального ряда. Ранговые переменные автоматически становятся качественными, если не учитывать упорядоченности их значений. Что касается качественных переменных, то каждому из возможных ее значений приходится сопоставлять дихотомную переменную, которая будет равна 1, если качественная переменная приняла данное значение, и 0 — в противном случае.

Отметим, что форма записи исходных данных, их сведение к одному типу, возможность использования только части данных и т. п., играют определенную роль при оценке практической эффективности вычислительного комплекса, предназначенного для решения задач классификации.

Матрица X не является единственным способом представления исходных данных в задачах кластерного анализа. Иногда исходная информация задана в виде **квадратной матрицы**

$$R=(r_{ij}),\ i,j=1,2,...,k,$$

элемент r_{ij} которой определяет степень близости i-го объекта к j-му.

Большинство алгоритмов кластерного анализа либо полностью исходит из матрицы расстояний (или близостей), либо требует вычисления отдельных ее элементов, поэтому, если данные представлены в форме *X,* то первым этапом решения задачи поиска кластеров будет выбор способа вычисления **расстояний или близости между объектами или признаками** (в этом отношении различие между объектами и признаками является существенным).

Относительно просто определяется близость между признаками. Как правило, кластерный анализ признаков преследует те же цели, что и факторный анализ — выделение групп связанных между собой признаков, отражающих определенную сторону изучаемых объектов. В этом случае мерами близости служат различные статистические коэффициенты связи.

Если признаки количественные, то можно использовать оценки обычных парных выборочных коэффициентов корреляции r_{ij}, i,j=1,2,...,k. Однако коэффициент корреляции измеряет только линейную связь, поэтому если связь нелинейна, то следует использовать корреляционное отношение, либо произвести подходящее преобразование шкалы признаков.

Существуют также различные коэффициенты связи, определенные для ранговых, качественных и дихотомных переменных.

3.1.2. Расстояние между объектами и мера близости

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов.

В общем случае **понятие однородности объектов** задается либо введением правила вычислений расстояния $\rho(X_i,X_j)$

между любой парой исследуемых объектов (X_1,X_2,..X_n), либо заданием некоторой функции r(X_i,X_j), характеризующей степень близости i-го и j-го объектов. Если задана функция ρ(X_i,X_j), то близкие с точки зрения этой метрики объекты считаются однородными, принадлежащими одному классу. При этом необходимо сопоставлять ρ(X_i,X_j) с некоторым пороговым значением, определяемым в каждом конкретном случае по-своему.

Аналогично используется и **мера близости** $r(X_i,X_j)$, при задании которой надо помнить о необходимости выполнения условий *симметрии* $r(X_i,X_j)=r(X_j,X_i)$; *максимального сходства объекта с самим собой* $r(X_i,X_j)=max$r(X_i,X_j), при $1 \le^{\wedge} j \le n$, и *монотонного убывания* r(X_i,X_j) по $\rho(X_i,X_j)$, т. е. j из $\rho(X_k,X_l) \ge \rho(X_i,X_j)$ должно следовать неравенство $r(X_k,X_l) < r(X_i,X_j)$.

Выбор метрики или меры близости является узловым моментом исследования, от которого в основном зависит окончательный вариант разбиения объектов на классы при данном алгоритме разбиения. В каждом конкретном случае этот выбор должен производиться по-своему в зависимости от целей исследования, физической и статистической природы вектора наблюдений X, априорных сведений о характере вероятностного распределения X.

Рассмотрим наиболее часто используемые расстояния и меры близости в задачах кластерного анализа.

РАССТОЯНИЕ МАХАЛАНОБИСА (ОБЩИЙ ВИД)

В случае зависимых компонент $x_1,x_2,...x_k$ вектора наблюдений X и их различной значимости в решении вопроса классификации обычно используют обобщенное (взвешенное) расстояние Махаланобиса, задаваемое формулой

$$\rho_0(X_i, X_j) = \sqrt{(X_i - X_j)^T \Lambda^T \sum^{-1} \Lambda (X_i - X_j)},$$

(135)

где Σ — ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения;

Λ — некоторая симметрическая неотрицательно-определенная матрица "весовых" коэффициентов, которая чаще всего выбирается диагональной.

Следующие три вида расстояний являются частными случаями метрики ρ_0 .

ОБЫЧНОЕ ЕВКЛИДОВО РАССТОЯНИЕ

$$\rho_E(X_i, X_j) = \sqrt{\sum_{l=1}^{k}(x_{il} - x_{jl})^2},$$

(136)

где x_{il}, x_{jl} — величина l-й компоненты у i-го (j-го) объекта(l=1,2,..., k; i,j=1,2,...,n).

Использование этого расстояния оправдано в случаях, если:

а) наблюдения берутся из генеральных совокупностей, имеющих многомерное нормальное распределение с ковариационной матрицей вида $\sigma^2 E_k$, т.е. компоненты X взаимно независимы и имеют одну и ту же дисперсию;

б) компоненты вектора наблюдений X однородны по физическому смыслу и одинаково важны для классификации;

в) признаковое пространство совпадает с геометрическим пространством.

Естественно с геометрической точки зрения и содержательной интерпретации евклидово расстояние может оказаться бессмысленным, если его признаки имеют разные единицы измерения. Для приведения признаков к одинаковым единицам прибегают к нормировке каждого признака путем деления центрированной величины на среднее квадратическое отклонение и переходят от матрицы X к нормированной матрице с элементами

$$x_{il}^{H} = \frac{x_{il} - \bar{x}_l}{s_l},$$

где x_{il} — значение l-го признака у i-го объекта;

$\bar{x}_l$ — среднее арифметическое значение l-го признака;

$s_l = \sqrt{\frac{1}{n}\sum(x_{il} - \bar{x}_l)^2}$ — среднее квадратическое отклонение l-го признака.

Однако эта операция может привести к нежелательным последствиям. Если кластеры хорошо разделены по одному

признаку и не разделены по другому, то после нормировки дискриминирующие возможности первого признака будут уменьшены в связи с увеличением "шумового" эффекта второго.

"ВЗВЕШЕННОЕ" ЕВКЛИДОВО РАССТОЯНИЕ

$$\rho_{BE}(X_i, X_j) = \sqrt{\sum_{l=1}^{k} w_l (x_{il} - x_{jl})^2} \qquad (137)$$

применяется в случаях, когда каждой компоненте x_l вектора наблюдений X удается приписать некоторый "вес" w_l, пропорциональный степени важности признака в задаче классификации. Обычно принимают $0 \le w_l \le 1$, где l=1,2,...,k.
Определение "весов", как правило, связано с дополнительными исследованиями, например организацией опроса экспертов и обработкой их мнений. Определение весов w_l только по данным выборки может привести к ложным выводам.

ХЕММИНГОВО РАССТОЯНИЕ

$$\rho_H(X_i, X_l) = \sum_{l=1}^{k} \left| x_{il} - x_{jl} \right| \qquad (138)$$

используется как мера различия объектов, задаваемых дихотомическими признаками. Хеммингово расстояние равно числу несовпадений значений соответствующих признаков в рассматриваемых i-м и j-м объектах.
В некоторых задачах классификации в качестве меры близости объектов можно использовать некоторые физически содержательные параметры, так или иначе характеризующие взаимоотношение между объектами. Например, задачу

классификации отраслей народного хозяйства с целью агрегирования решают на основе матрицы межотраслевого баланса.

В данной задаче объектом классификации является отрасль народного хозяйства, а матрица межотраслевого баланса представлена элементами Sij, характеризующими сумму годовых поставок i-й отрасли в j-ю в денежном выражении. В качестве меры близости $\{r_{ij}\}$ принимают **симметризованную нормированную матрицу межотраслевого баланса.** С целью нормирования денежное выражение поставок i-й отрасли в j-ю заменяют долей этих поставок по отношению ко всем поставкам i-й отрасли. Симметризацию нормированной матрицы межотраслевого баланса можно проводить выразив через среднее значение близость взаимных поставок между i-й и j-й отраслью так, что в этом случае $r_{ij}= r_{ji}$.

Как правило, решение задач классификации многомерных данных предусматривает в качестве предварительного этапа исследования реализацию методов, позволяющих выбрать из компонент $x_1,x_2,...,x_k$ наблюдаемых векторов x сравнительно небольшое число наиболее существенных информативных признаков, т. е. уменьшить размерность наблюдаемого пространства. С этой целью каждую из компонент $x_1,x_2,...,x_k$ рассматривают как объект, подлежащий классификации. После разбиения на небольшое число однородных в некотором смысле групп для дальнейшего исследования оставляют по одному представителю от каждой группы. При этом предполагается, что признаки, попавшие в одну группу, в определенном смысле связаны друг с другом и несут информацию о каком-то одном свойстве объекта.

В качестве близости между отдельными признаками обычно используют различные характеристики степени их коррелированности, в первую очередь коэффициенты корреляции. В ряде задач применяются и другие расстояния (метрики). Выбор метрики определяется структурой

признакового пространства и целью классификации. Формализовать этот этап задачи классификации пока не представляется возможным.

3.1.3. Расстояние между кластерами

В ряде процедур классификации (кластер-процедур) используют понятия **расстояния между группами объектов и меры близости двух групп объектов.**

Пусть S_i — i-я группа (класс, кластер), состоящая из n_i объектов;

$\overline{x}_i$ — среднее арифметическое векторных наблюдений S_i группы, т. е. "центр тяжести" i-й группы;

$\rho(S_l,S_m)$ — расстояние между группами S_l и S_m.

Наиболее употребительными расстояниями и мерами близости между классами объектов являются:

• расстояние, измеряемое по принципу "ближайшего соседа":

$$\rho_{\min}(S_l, S_m) = \min_{\substack{x_i \in S_l \\ x_j \in S_m}} \rho(x_i, x_j);$$

(139)

• расстояние, измеряемое по принципу "дальнего соседа":

$$\rho_{\min}(S_l, S_m) = \max_{\substack{x_i \in S_l \\ x_j \in S_m}} \rho(x_i, x_j); \qquad (140)$$

• расстояние, измеряемое по "центрам тяжести" групп:

$$\rho(S_l, S_m) = \rho(\bar{x}_l, \bar{x}_m); \qquad (141)$$

• расстояние, измеряемое по принципу "средней связи". Это расстояние определяется как среднее арифметическое всех попарных расстояний между представителями рассматриваемых групп

$$\rho_{cp}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} \rho(x_i, x_j). \qquad (142)$$

Академиком А. Н. Колмогоровым было предложено "обобщенное расстояние" между классами, которое в

качестве частных случаев включает в себя все рассмотренные выше виды расстояний.

Обобщенное расстояние основано на понятии так называемого "обобщенного среднего", а точнее — степенного среднего, и определяется формулой

$$\rho_{об}(S_l, S_m) = \left[\frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} \rho^r(x_i, x_j). \right]^{1/r} .$$

(143)

Можно показать, что при $r \to \infty$

$$\rho_{об}(S_l, S_m) = \rho_{\max}(S_l, S_m),$$

при $r \to -\infty$

$$\rho_{об}(S_l, S_m) = \rho_{\min}(S_l, S_m),$$

при $r = 1$

$$\rho_{об}(S_l, S_m) = \rho(S_l, S_m).$$

Из формулы (7.9) следует, что если $S_{(m,q)} = S_m \cup S_q$ — группа элементов, полученная путем объединения кластеров S_m и S_q, то обобщенное расстояние между кластерами S_l и $S_{(m,q)}$ определяется по формуле

$$\rho_{об}(S_l, S_{(m,q)} = \left\{ \frac{n_m \left[\rho_{об}(S_l, S_m)\right]^r + n_q \left[\rho_{об}(S_l, S_q)\right]^r}{n_m + n_q} \right\}^{1/r} .$$

(144)

Расстояние между группами элементов особенно важно в так называемых **агломеративных иерархических кластер-процедурах,** так как принцип работы таких алгоритмов состоит в последовательном объединении сначала самых близких элементов, а затем и целых групп все более и более отдаленных друг от друга элементов.

При этом расстояние между классами S_l и $S_{(m,q)}$, являющимися объединением двух других классов S_m и S_q, можно определить по формуле:

$$\rho_{l,(m,q)} = \rho(S_l, S_{(m,q)} = \alpha\rho_{lm} + \beta\rho_{lq} + \gamma\rho_{mq} + \delta \mid \rho_{lm} - \rho_{lq} \mid, \quad (145)$$

где $\rho_{lm} = \rho(S_l,S_m)$; $\rho_{lq} = \rho(S_l,S_q)$; $\rho_{mq} = \rho(S_m ,S_q)$ — расстояния между классами S_l, S_m и S_q;

α, β, γ и δ — числовые коэффициенты, значение которых определяет специфику процедуры, ее алгоритм.

Например, при $\alpha = \beta = -\delta = \frac{1}{2}$ и $\gamma = 0$ приходим к расстоянию, построенному по принципу "ближайшего соседа". При $\alpha = \beta = \delta = \frac{1}{2}$ и $\gamma = 0$ расстояние между классами определяется по принципу "дальнего соседа", как расстояние между двумя самыми дальними элементами этих классов. И наконец, при

$$\alpha = \frac{n_m}{n_m + n_q}, \quad \beta = \frac{n_q}{n_m + n_q}, \quad \gamma = \delta = 0$$

соотношение (145) приводит к расстоянию ρ между классами, вычисленному как среднее из расстояний между всеми парами элементов, один из которых берется из одного класса, а другой — из другого класса.

3.1.4. Функционалы качества разбиения

Существует большое количество различных способов разбиения на классы заданной совокупности элементов. Поэтому представляет интерес задача сравнительного анализа качества этих способов разбиения. С этой целью вводится понятие функционала качества разбиения Q(S), определенного на множестве всех возможных разбиений.

Наилучшее разбиение S* представляет собой такое разбиение, при котором достигается экстремум выбранного **функционала качества.** Следует отметить, что выбор того или иного функционала качества разбиения, как правило, опирается на эмпирические соображения.

Рассмотрим некоторые наиболее распространенные функционалы качества разбиения. Пусть исследованием выбрана метрика ρ в пространстве X и $S=(S_1,S_2,...,S_p)$ некоторое фиксированное разбиение наблюдений X_1, X_2,..., X_n на заданное число p классов $S_1,S_2,...,S_p$.

Существуют следующие характеристики функционала качества:

• сумма внутриклассовых дисперсий

$$Q_1(S) = \sum_{l=1}^{p} \sum_{x_i \in S_l} \rho^2(x_i \bar{x}_l); \qquad (146)$$

• сумма попарных внутриклассовых расстояний между элементами

$$Q_2(S) = \sum_{l=1}^{p} \sum_{x_i, x_j \in S_l} \rho^2(x_i x_j), \qquad (147)$$

или

$$Q_2'(S) = \sum_{l=1}^{p} \frac{1}{n_l} \sum_{x_i, x_j \in S_l} \rho^2(x_i x_j).$$

$Q_1(S)$ и $Q_2(S)$ широко используются в задачах кластерного анализа для сравнения качества процедур разбиения;

• обобщенная внутриклассовая дисперсия

$$Q_3(S) = \det(\sum_{l=1}^{p} n_l W_l), \qquad (148)$$

где det A — определитель матрицы A;

W_l — выборочная ковариационная матрица класса S_l, элементы которой определяются по формуле

$$w_{qm}(l) = \frac{1}{n_l} \sum_{x_i \in S_l} (x_{iq} - \bar{x}_q)(x_{im} - \bar{x}_m), \quad q, m = 1, 2, \ldots, k,$$

где x_{iq} — q-я компонента многомерного наблюдения x_i;
$\bar{x}_q$ — среднее значение q-й компоненты, вычисленное по наблюдениям *l*-го класса.

Качество разбиения характеризуют и другим видом обобщенной дисперсии, в которой операция суммирования W_l заменена операцией умножения

$$Q_4(S) = \prod_{l=1}^{p} (\det W_l)^{n_l}.$$

Отметим, что функционалы $Q_3(S)$ и $Q_4(S)$ обычно используют при решении вопроса: не сосредоточены ли наблюдения, разбитые на классы, в пространстве размерности, меньшей, чем *k*.

3.1.5. Иерархические кластер-процедуры

Иерархические (деревообразные) процедуры являются наиболее распространенными алгоритмами кластерного анализа по их реализации на ЭВМ. Они бывают двух типов: агломеративные и дивизимные. В **агломеративных** процедурах начальным является разбиение, состоящее из *n* одноэлементных классов, а конечным — из одного класса; в **дивизимных** наоборот.

Принцип работы иерархических агломеративных (дивизимных) процедур состоит в последовательном объединении (разделении) групп элементов сначала самых близких (далеких), а затем все более отдаленных (близких) друг от друга. Большинство этих алгоритмов исходит из матрицы расстояний (сходства).

К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации. Алгоритмы требуют на каждом шаге матрицы вычисления расстояний, а следовательно, емкой машинной памяти и большого количества времени. В этой связи реализация таких алгоритмов при числе наблюдений, большем нескольких сотен, нецелесообразна, а в ряде случаев и невозможна.

Приведем пример агломеративного иерархического алгоритма. На первом шаге каждое наблюдение X_i (i=1,2,..., п) рассматривается как отдельный кластер. В дальнейшем на каждом шаге работы алгоритма происходит объединение самых близких кластеров, и, с учетом принятого расстояния, по формуле пересчитывается матрица расстояний, размерность которой, очевидно, снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс. Большинство программ, реализующих алгоритм иерархической классификации, предусматривают графическое представление классификации в виде дендрограммы.

Пример

Провести классификацию $n=6$ объектов, каждый их которых характеризуется двумя признаками:

№ объекта i	1	2	3	4	5	6
x_{i1}	5	6	5	10	11	10
x_{i2}	10	12	13	9	9	7

Расположение объектов в виде точек на плоскости показано на рис. 15.

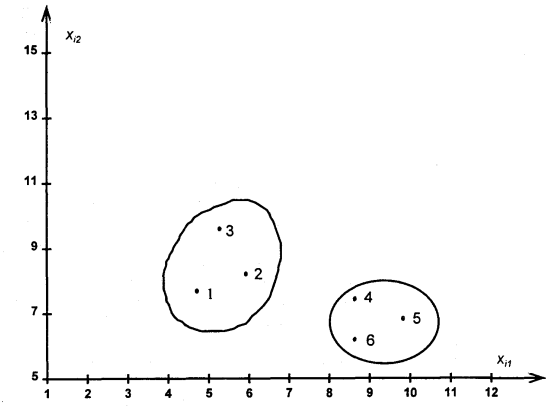


Рис. 15. Классификация объектов

Решение

Воспользуемся агломеративным иерархическим алгоритмом классификации. В качестве расстояния между объектами возьмем обычное евклидово расстояние. Тогда согласно формуле (136) расстояние между первым и вторым объектами

$$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = 2{,}24,$$

а между первым и третьим объектами

$$\rho_{13} = \sqrt{(5-5)^2 + (10-13)^2} = 3.$$

Очевидно, что

$$\rho_{11} = 0.$$

Аналогично находим расстояния между шестью объектами и строим матрицу расстояний.

$$R_1 = \left\{\rho(x_i, x_j)\right\} = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 6{,}08 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 5{,}83 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}21 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 1 & 2 \\ 6{,}08 & 5{,}83 & 7{,}21 & 1 & 0 & 2{,}24 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2 & 2{,}24 & 0 \end{pmatrix}.$$

Из матрицы расстояний следует, что четвертый и пятый объекты наиболее близки $\rho_{4,5} = 1{,}00$ и поэтому объединяются в один кластер. После объединения объектов имеем пять кластеров:

Номер кластера	1	2	3	4	5
Состав кластера	(1)	(2)	(3)	(4,5)	(6)

Расстояние между кластерами определим по принципу "ближайшего соседа", воспользовавшись формулой пересчета (145). Так расстояние между объектом S_1 и кластером $S_{(4,5)}$

$$\rho_{1,(4,5)} = \rho(S_1, S_{(4,5)} = \frac{1}{2}\rho_{14} + \frac{1}{2}\rho_{15} - \frac{1}{2}|\,\rho_{14} - \rho_{15}\,| =$$

$$= \frac{1}{2}[5{,}10 + 6{,}08] - \frac{1}{2}[|\,5{,}10 - 6{,}08\,|] = 5{,}10.$$

Таким образом, расстояние $\rho_{1,(4,5)}$ равно расстоянию от объекта 1 до ближайшего к нему объекта, входящего в кластер S(4,5) , т.е. $\rho_{1,(4,5)} = \rho_{1,4} = 5{,}10$. Тогда матрица расстояний:

$$R_2 = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 2 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2 & 0 \end{pmatrix}.$$

Объединим второй и третий объекты, имеющие наименьшее расстояние $\rho_{2,3} = 1{,}41$. После объединения объектов имеем четыре кластера:

$$S_{(1)}, S_{(2,3)}, S_{(4,5)}, S_{(6)}.$$

Вновь найдем матрицу расстояний. Для того чтобы рассчитать расстояние до кластера $S_{2,3}$, воспользуемся

матрицей расстояний R_2. Например, расстояние между кластерами $S_{(4,5)}$ и $S_{(2,3)}$ равно

$$\rho_{(4,5),(2,3)} = \frac{1}{2}\rho_{(4,5),2} + \frac{1}{2}\rho_{(4,5),3} - \frac{1}{2}|\rho_{4,5),2} - \rho_{(4,5),3}| =$$

$$= \frac{5}{2} + \frac{6{,}40}{2} - \frac{1{,}40}{2} = 5.$$

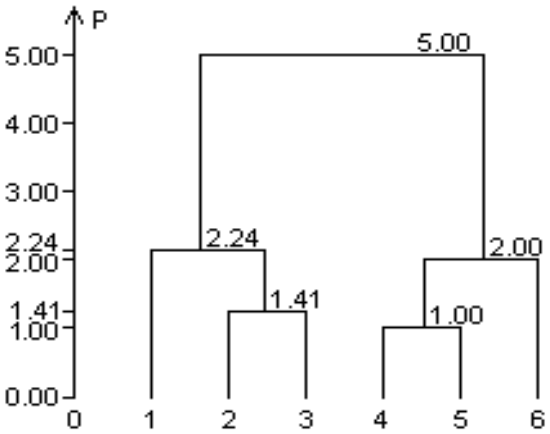


Рис. 16**.** Дендрограмма

Проведя аналогичные расчеты, получим

$$R_3 = \begin{pmatrix} 0 & 2{,}24 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 5 & 6{,}40 \\ 5{,}10 & 5 & 0 & 2 \\ 5{,}83 & 6{,}40 & 2 & 0 \end{pmatrix}.$$

Объединим кластеры $S_{(4,5)}$ и $S_{(6)}$, расстояние между которыми, согласно матрице R_3, наименьшее $\rho_{(4,5),6} = 2$. В результате получим три кластера:

$$S_{(1)},\ S_{(2,3)}\ \text{и}\ S_{(4,5,6)}.$$

Матрица расстояний будет иметь вид:

$$R_4 = \begin{pmatrix} 0 & 2{,}24 & 5{,}10 \\ 2{,}24 & 0 & 5 \\ 5{,}10 & 5 & 0 \end{pmatrix}.$$

Объединим теперь кластеры $S_{(1)}$ и $S_{(2,3)}$, расстояние между которыми $\rho_{1,(2,3)} = 2{,}24$. В результате получим два кластера: $S_{(1,2,3)}$ и $S_{(4,5,6)}$, расстояние между которыми, найденное по принципу "ближайшего соседа", $\rho_{(1,2,3),\,(4,5,6)} = 5$.

Результаты иерархической классификации объектов представлены на рис. 16 в виде дендрограммы.

На рис.16 приводятся расстояния между объединяемыми на данном этапе кластерами (объектами). В нашем примере предпочтение следует отдать предпоследнему этапу классификации, когда все объекты объединены в два кластера (рис. 16):

$S_{(1,2,3)}$ и $S_{(4,5,6)}$.

3.2. Дискриминантный анализ

3.2.1. Методы классификации с обучением

Дискриминантный анализ как раздел **многомерного статистического анализа** включает в себя статистические методы классификации многомерных наблюдений в ситуации, когда исследователь обладает так называемыми обучающими выборками ("классификация с учителем").

В общем случае задача различения (дискриминации) формулируется следующим образом. Пусть результатом наблюдения над объектом является реализация *k*-мерного случайного вектора $x = (x_1,x_2,...,x_k)^T$. Требуется установить правило, согласно которому по наблюденному значению вектора x объект относят к одной из возможных совокупностей π_1, $i = 1,2,...,l$. Для построения правила дискриминации все выборочное пространство R значений вектора x разбивается на области R_i , $i=1,2,...,l$, так, что при попадании x в R_i объект относят к совокупности π_i.

Правило дискриминации выбирается в соответствии с определенным принципом оптимальности на основе априорной информации о совокупностях ρ_i, извлечения объекта из π_i. При этом следует учитывать размер убытка от неправильной дискриминации. Априорная информация может быть представлена как в виде некоторых сведений о функции *k*-мерного распределения признаков в каждой совокупности, так и в виде выборок из этих совокупностей. Априорные вероятности ρ_i могут быть либо заданы, либо нет. Очевидно, что рекомендации будут тем точнее, чем полнее исходная информация.

С точки зрения применения дискриминантного анализа наиболее важной является ситуация, когда исходная информация о распределении представлена выборками из

них. В этом случае задача дискриминации ставится следующим образом.

Пусть $x_1^{(i)},...,x_j^{(i)},...,x_{ni}^{(i)}$ — выборка из совокупности π_i, $i=1,2,...,l$; $j=1,2,...,$ n_i, причем каждый j-й объект выборки представлен k-мерным вектором параметров $x_j^{(i)} = (x_{j1}^{(i)}...x_{iq}^{(i)}...x_{ik}^{(i)})^T$. Произведено дополнительное наблюдение $x= (x_1,...,x_k)^T$ над объектом, принадлежащим одной из совокупностей π_i. Требуется построить правило отнесения наблюдения x к одной из этих совокупностей.

Обычно в задаче различения переходят от вектора признаков, характеризующих объект, к линейной функции от них, дискриминантной функции — гиперплоскости, наилучшим образом разделяющей совокупность выборочных точек.

Наиболее изучен случай, когда известно, что распределение векторов признаков в каждой совокупности нормально, но нет информации о параметрах этих распределений. Здесь естественно заменить неизвестные параметры распределения в дискриминантной функции их наилучшими оценками. Правило дискриминации можно основывать на отношении правдоподобия.

Непараметрические методы дискриминации не требуют знаний о точном функциональном виде распределений и позволяют решать задачи дискриминации на основе незначительной априорной информации о совокупностях, что особенно ценно для практических применений.

В параметрических методах эти точки используются для оценки параметров статистических функций распределения. В параметрических методах построения функции, как правило, используется нормальное распределение.

3.2.2. Линейный дискриминантный анализ

Предположения:

1) имеются разные классы объектов;
2) каждый класс имеет нормальную функцию плотности от k переменных:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_{ni} \end{pmatrix};$$

$$f_i(x) = (2\pi)^{ni/2} \left|\Sigma_i\right|^{-1/2} \exp\left(-\frac{1}{2}\left(x - \mu^{(i)}\right)^T \sum\nolimits_i^{-1} \left(x - \mu^{(i)}\right)\right),$$

(149)

где $\mu^{(i)}$ — вектор математических ожиданий переменных размерности k;
Σ_i — ковариационная матрица при $n=n_i$,
Σ_i^{-1}, — обратная матрица.
Матрица Σ_i — положительно определена.
В случае, если параметры известны дискриминацию можно провести следующим образом.
Имеются функции плотности $f_1(x),f_2(x),\dots,f_l(x)$ нормально распределенных классов. Задана точка x в пространстве k измерений. Предполагая, что $f_i(x)$ имеет наибольшую плотность, отнесем точку x к i-му классу. Существует доказательство, что если априорные вероятности для определяемых точек каждого класса одинаковы и потери при неправильной классификации i-й группы в качестве j-й не зависят от i и j, то решающая процедура минимизирует ожидаемые потери при неправильной классификации.
Приведем пример оценки параметра многомерного нормального распределения μ и Σ.

μ и Σ могут быть оценены по выборочным данным $\hat{\mu}$ и $\hat{\Sigma}$ для классов. Задано l выборок $(x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}) \equiv x_i; (i = \overline{1,l})$ из некоторых классов. Математические ожидания μ_1, $\mu_2, \dots$, μ_k могут быть оценены средними значениями

$$\hat{\mu}_q^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{jq}^{(i)}; q = \overline{1,k}. \qquad (150)$$

Несмещенные оценки элементов ковариационной матрицы Σ есть

$$(\hat{\Sigma}_i)_{rs} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{jr}^{(i)} - \hat{\mu}_r{}^{(i)})(x_{jr}^{(i)} - \hat{\mu}_r{}^{(i)}); \ \mathrm{r,s} = \overline{1,k}. \qquad (151)$$

Следовательно, можно определить $\hat{\mu}^{(i)}$ и $\hat{\Sigma}_i$ по l выборкам в каждом классе при помощи (150), (151). Получив оценки, точку x отнесем к классу, для которой функция $f(x)$ максимальна.

Введем предположение, что все классы, среди которых должна проводиться дискриминация, имеют нормальное распределение с одной и той же ковариационной матрицей Σ.

В результате существенно упрощается выражение для дискриминантной функции.

Класс, к которому должна принадлежать точка $\boldsymbol{x}$, можно определить на основе неравенства

$$f_i(x) > f_j(x). \qquad (152)$$

Воспользуемся формулой (149) для случая, когда их ковариационные матрицы равны: $\Sigma_i = \Sigma_j = \Sigma$, а $\mu^{(i)}$ есть вектор математических ожиданий класса i. Тогда (152) можно представить неравенством их квадратичных форм:

$$-[(x-\mu^{(i)})^T \Sigma^{-1}(x-\mu^{(i)})] > -[(x-\mu^{(j)})^T \Sigma^{-1}(x-\mu^{(j)})].$$

Раскроем скобки:

$$-(x^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu^{(i)} - \mu^{(i)^T} \Sigma^{-1} x + \mu^{(i)^T} \Sigma^{-1} \mu^{(j)}) >$$

$$> -(x^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu^{(j)} - \mu^{(j)^T} \Sigma^{-1} x + \mu^{(j)^T} \Sigma^{-1} \mu^{(j)}).$$

(153)

Вспомним, если имеем два вектора Z и W, то скалярное произведение можно записать $Z^T W = W^T Z = (Z,W)$. В выражении (153) исключим $x^T \Sigma^{-1} x$ справа и слева, поменяем у всех членов суммы знаки. Теперь преобразуем:

$$x^T \Sigma^{-1} \mu^{(j)} + \mu^{(j)^T} \Sigma^{-1} x = 2(x^T \Sigma^{-1} \mu^{(j)}) = 2(x, \Sigma^{-1} \mu^{(j)});$$

$$\mu^{(j)^T} \Sigma^{-1} \mu^{(j)} = (\mu^{(j)}, \Sigma^{-1} \mu^{(j)}).$$

Аналогично провести преобразования по индексу i представим читателю самостоятельно.

Сократим правую и левую часть неравенства (153) на 2 и, используя запись квадратичных форм, получим

$$(x, \Sigma^{-1} \mu^{(i)}) - \frac{1}{2}(\mu^{(i)}, \Sigma^{-1} \mu^{(i)}) > (x, \Sigma^{-1} \mu^{(j)}) - \frac{1}{2}(\mu^{(j)}, \Sigma^{-1} \mu^{(j)}).$$

(154)

Введем обозначения в выражение (154):

$$v^{(i)} = \Sigma^{-1} \mu^{(i)}; \; i = \overline{1,m};$$

$$\lambda_i = \frac{1}{2}(\mu^{(i)}, \Sigma^{-1} \mu^{(i)}); \quad i = \overline{1,m}.$$

Тогда выражение (154) примет вид

$$(x, v^{(i)}) - \lambda_i > (x, v^{(j)}) - \lambda_j. \qquad (155)$$

Следствие: проверяемая точка x относится к классу i, для которого линейная функция

$$h_i(x) = (x, v^{(i)}) - \lambda_i = \max. \qquad (156)$$

Преимущество метода линейной дискриминации Фишера заключается в линейности дискриминантной функции (156) и надежности оценок ковариационных матриц классов.

Пример

Имеются два класса с параметрами $(\mu^{(1)},\Sigma_1)$ и $(\mu^{(2)},\Sigma_2)$. По выборкам из этих совокупностей объемом n_1 и n_2 получены оценки $\hat{\Sigma}_1$ и $\hat{\Sigma}_2$. Первоначально проверяется гипотеза о том, что ковариационные матрицы Σ_1 и Σ_2 равны. В случае, если оценки $\hat{\Sigma}_1$ и $\hat{\Sigma}_2$ статистически неразличимы, то принимается, что $\Sigma_1=\Sigma_2=\Sigma$ и строится общая оценка $\hat{\Sigma}$, основанная на суммарной выборке объемом n_1+n_2, после чего строится линейная дискриминантная функция Фишера (156). Существуют и другие методы. Так, в математическом обеспечении пакета "Олимп" используется пошаговый дискриминантный анализ.

3.2.3. Дискриминантный анализ при нормальном законе распределения показателей

Имеются две генеральные совокупности X и У, имеющие трехмерный нормальный закон распределения с неизвестными, но равными ковариационными матрицами. Из них взяты обучающие выборки с объемами n_1уX и n_2уY.

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \dots & \dots & \dots \\ x_{n_1 1} & x_{n_1 2} & x_{n_1 3} \end{pmatrix}, \qquad (157)$$

$$Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \dots & \dots & \dots \\ y_{n_2 1} & y_{n_2 2} & y_{n_2 3} \end{pmatrix}. \qquad (158)$$

Целью дискриминантного анализа является отнесение нового наблюдения (строки матрицы *Z*) либо к *X*, либо к Y.

$$Z = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ \dots & \dots & \dots \\ z_{l1} & z_{l2} & z_{l3} \end{pmatrix}. \qquad (159)$$

Для решения задачи по обучающим выборкам определим векторы средних:

$$\overline{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} \text{ и } \overline{Y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix}.$$

1. Определим оценки ковариационных матриц:

$$S_x = (s_{ki})_x \text{ и } S_y = (s_{ki})_y;\ \bar{x}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}.$$

Найдем элемент матрицы S_x:

$$s_{kj}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = \overline{x_j x_k} - \bar{x}_j \bar{x}_k;\ k = 1,2,3,$$

где $\bar{x}_j$ и $\bar{x}_k$ — средние значения.

2. Рассчитаем несмещенную оценку суммарной ковариационной матрицы:

$$\hat{S} = \frac{1}{n_1 + n_2 - 2}\left(n_1 S_x + n_2 S_y\right).$$

3. Определим матрицу $\hat{S}^{-1}$, обратную к $\hat{S}$.

4. Вычислим вектор оценок коэффициентов дискриминантной функции $a = \hat{S}^{-1}(\overline{X} - \overline{Y})$.

5. Рассчитаем оценки векторов значений дискриминантной функции для матриц исходных данных $\hat{U}_x = Xa,\ \hat{U}_y = Ya$.

6. Вычислим средние значения оценок дискриминантной функции:

$$\overline{\hat{u}}_x = \frac{1}{n_1}\sum_{i=1}^{n_1}\hat{u}_{xi},\quad \overline{\hat{u}}_y = \frac{1}{n_2}\sum_{i=1}^{n_2}\hat{u}_{yi}.$$

7. Определим константу $\hat{C} = \frac{1}{2}(\overline{\hat{u}}_x + \overline{\hat{u}}_y)$.

Дискриминантную функцию для v-ого наблюдения, подлежащего дискриминации, получим, решив уравнение

$$\hat{u}_v = z_{v1} + a_1 + z_{v2}a_2 + z_{v3}a_3.$$

Если $\hat{u}_v \geq \hat{C}$, то v-е наблюдение надо отнести к совокупности x, если же $\hat{u}_v < \hat{C}$, то v -е наблюдение следует отнести к совокупности y.