Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Комсомольский-на-Амуре государственный технический университет»

**Г. М. Гринфельд, А. В. Моисеев**

# МЕТОДЫ ОПТИМИЗАЦИИ ЭКСПЕРИМЕНТА В ХИМИЧЕСКОЙ ТЕХНОЛОГИИ

Утверждено в качестве учебного пособия
Ученым советом Федерального государственного бюджетного
образовательного учреждения высшего профессионального образования
«Комсомольский-на-Амуре государственный технический университет»

Комсомольск-на-Амуре
2014

ББК 22.172:35я7
УДК 66:51-7(07)
Г854

*Рецензенты:*
Кафедра «Технология нефте- и углехимических производств»
ФГБОУ ВПО «Санкт-Петербургский государственный
технологический институт (технический университет)»,
зав. кафедрой доктор химических наук, профессор В. М. Потехин;
А. А. Кулик, кандидат технических наук,
начальник отдела организации разработки и проведения экспертизы ПСД
ООО «РН – Комсомольский НПЗ»

**Гринфельд, Г. М.**

Г854 Методы оптимизации эксперимента в химической технологии : конспект лекций / Г. М. Гринфельд, А. В. Моисеев. – Комсомольск-на-Амуре : ФГБОУ ВПО «КнАГТУ», 2014. – 76 с.
ISBN 978-5-7765-1102-8

Рассматриваются вопросы построения стохастических моделей химико-технологических объектов с использованием методов корреляционного и регрессионного анализов, принципы построения оптимальных планов проведения экспериментов, обеспечивающих сокращение количества проводимых на объекте экспериментов при заданном уровне адекватности полученных моделей.

Конспект лекций предназначен для студентов, обучающихся по направлению подготовки «Химическая технология и биотехнология» всех форм обучения.

ББК 22.172:35я7
УДК 66:51-7(07)

ISBN 978-5-7765-1102-8

© ФГБОУ ВПО «Комсомольский-на-Амуре государственный технический университет», 2014

## ОГЛАВЛЕНИЕ

# ВВЕДЕНИЕ

Актуальность данного пособия определяется возрастающими потребностями химической промышленности в специалистах, обладающих достаточной квалификацией для управления современными химико-технологическими объектами, состоящими из большого числа аппаратов и связей (потоков) между ними.

Зачастую только результаты экспериментальных исследований, проводимых на таких объектах, являются той информацией, на основании обработки которой по полученной математической модели объекта может быть реализован качественный алгоритм управления данным объектом.

В то же время построение математической модели, описывающей технологический процесс с необходимой точностью, зачастую сопряжено с огромным объемом работы на стадии проведения экспериментов.

Задача планирования экспериментов состоит в установлении минимально необходимого их количества и условий их проведения, в выборе методов математической обработки результатов и в принятии решений. Планирование экспериментов значительно сокращает их количество, необходимое для получения модели процесса.

## 1. ДЕТЕРМИНИРОВАННЫЕ И СТОХАСТИЧЕСКИЕ МОДЕЛИ ХИМИКО-ТЕХНОЛОГИЧЕСКИХ ОБЪЕКТОВ

Современные промышленные химико-технологические объекты (ХТО) представляют собой сложные многостадийные системы с высокой стоимостью конечного продукта, большим числом взаимосвязанных переменных, подверженные существенному влиянию неконтролируемых воздействий. Характерными для таких объектов являются недостаточная теоретическая изученность протекающих в них процессов и связанная с этим невозможность выбора оптимального режима.

В связи с этим возникает необходимость в разработке некоторой модели объекта, по которой можно было бы всесторонне исследовать различные режимы работы, в том числе те, которые по условиям безопасности или экономической целесообразности не могут быть осуществлены на реальном объекте.

Среди большого числа признаков, по которым может быть осуществлена классификация моделей ХТО, рассмотрим их деление на ***детерминированные*** и ***стохастические*** модели.

Если режим функционирования ХТО достаточно хорошо изучен и имеется полная информация обо всех происходящих в нем физико-химических процессах, то можно построить ***детерминированную модель объекта***. При этом необходимо основываться на уравнениях всех протекаю-

щих в объекте химических реакций, учитывать гидродинамические режимы перемещения реагентов, скорости диффузии и теплопередачи, материальный и тепловой балансы, фазовые превращения. Естественно, что такая модель будет в полной мере соответствовать ХТО (будет адекватна ему), но с высокой степенью вероятности она окажется громоздкой и не пригодной для оперативного управления и оптимизации ХТО.

В отличие от детерминированной модели, ***стохастическая модель*** формируется не на теоретических представлениях о совокупности всех протекающих в ХТО процессов, а на основании результатов обработки данных эксперимента, проведенного на этом объекте.

На рис. 1.1. приведена обобщенная структура ХТО [1].

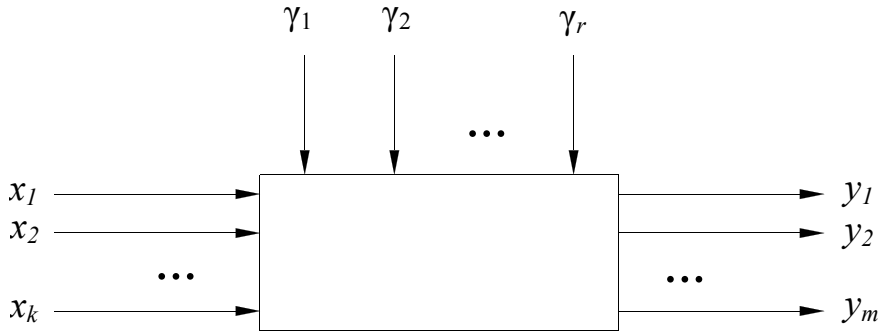


Рис. 1.1. Обобщенная структура ХТО

В приведенной структуре ХТО $x_i$ ($i$=1, …, $k$) – различные управляемые или контролируемые переменные (***входные или независимые переменные, факторы***), определяющие технологический режим процесса. Ими могут быть расход, концентрация и продолжительность взаимодействия реагентов, состав исходной реакционной смеси, температура и давление, содержание примеси и т. д. Факторы должны быть управляемы, т. е. они могут быть установлены на определенный уровень, который будет оставаться неизменным в течение всего опыта.

Варьирование уровней факторов от опыта к опыту приводит к различным изменениям отражающих качество функционирования ХТО показателей $y_i$ ($i$=1, …, $m$), определяемых как независимые или выходные переменные (***функции отклика***). К их числу могут относиться выход продукта, скорость кристаллизации, степень разложения, оптическая плотность, растворимость и т. д. Функции отклика могут определять и технико-экономические показатели, например, производительность установки или выход продукта.

Стохастическая модель ХТО представляет собой математическую зависимость, связывающую выходные и входные переменные:

$$\hat{y}_i = f_i(x_1, x_2, \dots, x_k).$$

Основным требованием, предъявляемым к такой модели, является ее адекватность объекту, на котором проводится эксперимент. При выполнении этого требования различие между значением выходной переменной, полученным при проведении эксперимента на объекте с определенными значениями входных переменных, и величиной $\hat{y}$, рассчитанной на модели при тех же значениях факторов, должно быть незначительным. Одна из причин неадекватности модели связана с воздействием на ХТО ряда неконтролируемых, случайным образом изменяющихся факторов $\gamma_i$ $(i = 1, \dots, r)$, таких как падение активности катализатора, изменение состояния поверхности теплообменной аппаратуры, присутствие невыявленных примесей. Влияние неучтенных факторов на величину функции отклика определяет случайный характер ее изменения. Это обуславливает необходимость использования статистических методов для обработки и анализа экспериментальных данных.

## 1.1. Пассивный и активный подходы к проведению эксперимента

Существуют два основных вида подходов к проведению эксперимента на ХТО: ***пассивный*** и ***активный***.

При пассивном эксперименте ХТО находится в режиме нормальной эксплуатации (в штатном режиме), а сам эксперимент состоит только в регистрации значений переменных в ходе проводимых на объекте опытов. Необходимо учитывать, что в условиях нормальной эксплуатации изменения выходных и входных переменных могут быть не достаточными для того, чтобы выявить все существенные особенности объекта, т. к. на уровне помех они могут быть незначительными. Но при этом затраты на проведение пассивного эксперимента обычно невелики.

В ходе активного эксперимента производятся целенаправленные, заранее спланированные воздействия на ХТО. В этом случае ***план эксперимента*** содержит информацию о значениях факторов в каждом из опытов, включенных в данный план, а также определяет порядок их выполнения. Следует отметить, что при проведении активного эксперимента возможен срыв технологического режима или поломка оборудования, но при этом существенно сокращается объем работы как на стадии проведения эксперимента, так и при обработке его результатов, а полученные модели ХТО обычно более надежны и достоверны, чем при пассивном эксперименте.

# 2. ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

## 2.1. Непрерывные и дискретные случайные переменные

Решение задачи планирования эксперимента и обработки полученных данных предполагает владение математическими приемами и методами, изучаемыми в курсе «Теория вероятности и математическая статистика». В данном разделе приводятся лишь основные понятия этого курса.

В процессе контроля значений переменных при проведении эксперимента возможны ***два случая***.

***Случай 1:*** Переменная $x_i$ ХТО принимает одно из нескольких возможных значений $(x_{i1}, x_{i2}, \dots, x_{in})$. Такие переменные относятся к дискретным, при этом регистрация переменной на уровне $x_{ij}$ производится с вероятность $p_j$. Очевидно, что

$$\sum_{j=1}^{n} p_j = 1.$$

Соответствие между всеми возможными значениями дискретной случайной величины и их вероятностями называют ***рядом распределения*** или ***вероятностным рядом:***

| | | | | | |
|---|---|---|---|---|---|
| $x_{i1}$ | $x_{i2}$ | …. | $x_{ij}$ | …. | $x_{in}$ |
| $p_1$ | $p_2$ | …. | $p_j$ | …. | $p_n$ |

***Случай 2:*** Переменная принимает произвольные значения внутри некоторого интервала, определенного технологическим регламентом. В этом случае переменные относятся к ***непрерывным случайным переменным***.

## 2.2. Свойства функции распределения и плотности распределения случайной переменной

Как непрерывные, так и дискретные случайные переменные в полной мере описываются с помощью ***функции распределения*** $\boldsymbol{F(x)}$ (интегральной функции распределения). Вид функции распределения, как и рассмотренный выше ряд распределения, являются различными формами задания ***закона распределения*** случайной величины.

Функцией распределения называют функцию $F(x)$, определяющую для каждого неслучайного значения $x$ вероятность $P$ того, что случайная величина $X$ примет значение меньше $x$:

$$F(x) = P(X < x).$$

Функция распределения $F(x)$ обладает следующими свойствами:

1) Областью значений функции распределения является отрезок [0; 1], т. е. $0 \leq F(x) \leq 1$.

2) Функция распределения – неубывающая функция, т. е. если $a \leq b$, то $F(a) \leq F(b)$.

3) Вероятность того, что случайная величина $X$ примет значение, заключенное в интервале $(a; b)$, равна приращению функции распределения на этом интервале:

$$P(a < X < b) = F(b) - F(a). \qquad (2.1)$$

4) Для функции распределения справедливы следующие предельные отношения:

$$\lim_{x \to -\infty} F(x) = 0,$$

$$\lim_{x \to \infty} F(x) = 1.$$

5) Для дискретной случайной переменной $x_i$, которая может принимать значения $x_{i1}, x_{i2}, \dots, x_{in}$, функция распределения имеет вид

$$F(x) = \sum_{x_{ij} < x} p_j,$$

где неравенство под знаком суммы указывает, что суммируются вероятности $p_j$ всех значений $x_{ij}$, величина которых меньше $x$.

Закон распределения непрерывной случайной переменной ХТО, имеющей непрерывную и дифференцируемую функцию распределения $F(x)$, чаще всего задается в виде ***плотности распределения*** $f(x)$, которая равна производной от функции $F(x)$:

$$f(x) = \frac{dF(x)}{dx}.$$

Функцию $f(x)$ называют также дифференциальной функцией распределения.

Функция плотности распределения $f(x)$ обладает следующими свойствами:

1) Так как плотность распределения $f(x)$ является производной неубывающей функции $F(x)$, то она неотрицательна: $f(x) \geq 0$.

2) Так как функция $F(x)$ – первообразная для $f(x)$, то

$$\int_a^b f(x)dx = F(b) - F(a),$$

следовательно,

$$P(a < X < b) = \int_a^b f(x)dx. \qquad (2.2)$$

3) Если в формуле (2.1) положить $a = -\infty$ и $b = \infty$, то получим

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

т. к. $P(-\infty < X < \infty)$ – вероятность достоверного события.

4) Выражение для функции распределения можно получить, полагая в формуле (2.2) $a = -\infty$ и $b = x$:

$$F(x) = \int_{-\infty}^{x} f(x)dx. \qquad (2.3)$$

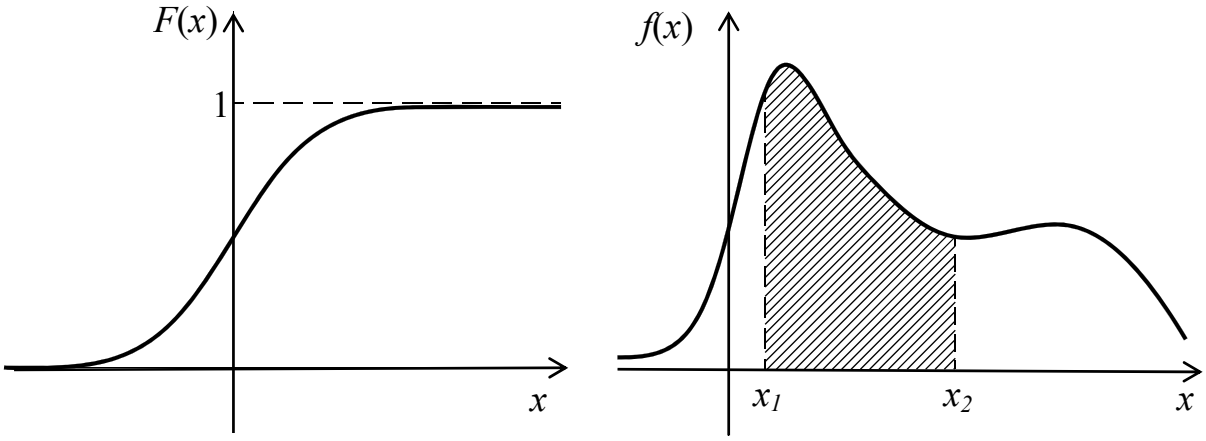


Рис. 2.1. Графики функции распределения $F(x)$
и плотности распределения $f(x)$ непрерывной случайной переменной

На рис. 2.1. приведены графики функции распределения $F(x)$ и плотности распределения $f(x)$ непрерывной случайной переменной.

## 2.3. Числовые характеристики законов распределения

Функция распределения и плотность распределения содержат полную информацию о случайной величине, но их определение не всегда достижимо. Для решения прикладных задач планирования эксперимента зачастую достаточно информации, задаваемой ***числовыми характеристиками случайной величины***.

К числовым характеристикам случайной величины относятся:

1) числовые характеристики, определяющие положение функции распределения случайной величины на числовой оси (***математическое ожидание*** $M(X)$, ***мода*** $Mo$, ***медиана*** $Me$, ***квантили распределения*** $x_p$);

2) числовые характеристики, зависящие от разброса случайной величины около значения математического ожидания (***дисперсия*** $D(X)$, ***среднее квадратическое отклонение*** $\sigma(x)$, ***коэффициент вариации*** $V(X)$);

3) числовые характеристики, определяемые по форме графика функции плотности распределения $f(x)$ (***коэффициент асимметрии*** $As$ и ***коэффициент эксцесса*** $Ex$).

Рассмотрим подробнее каждую из указанных характеристик.

***Математическое ожидание*** случайной величины $X$ указывает некоторое среднее значение, около которого группируются все возможные значения $X$. Для дискретной случайной переменной $X$, которая в ходе эксперимента принимает только одно из $n$ возможных значений $x_i$ с вероятностью $p_i$, математическое ожидание $M(X)$ равно:

$$M(X) = m_x = \sum_{i=1}^{n} x_i \, p_i.$$

Для непрерывной случайной величины $X$, имеющей заданную плотность распределения $f(x)$, математическое ожидание $M(X)$ равно:

$$M(X) = m_x = \int_{-\infty}^{\infty} x f(x) dx.$$

В «Курсе теории вероятностей» [2] рассматривались начальные моменты $m_k$ и центральные моменты $\mu_k$ непрерывной случайной величины -го порядка:

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

$$\mu_k = \int_{-\infty}^{\infty} (x - m_x)^k f(x) dx.$$

Очевидно, что математическое ожидание является начальным моментом 1-го порядка.

Математическое ожидание $M(X)$ обладает следующими свойствами:

1) $M(C) = C$, где $C$ = const;
2) $M(C \cdot X) = C \cdot M(X)$;
3) для произвольных случайных величин $X$ и $Y$ справедливо равенство

$$M(X \pm Y) = M(X) \pm M(Y);$$

4) для независимых случайных величин $X$ и $Y$ справедливо равенство

$$M(X \cdot Y) = M(X) \cdot M(Y).$$

Две случайные величины называются ***независимыми***, если закон распределения одной из них не зависит от того, какие возможные значения принимает другая величина.

***Модой*** дискретной случайной величины, обозначаемой $Mo$, называется ее значение, реализуемое в ходе эксперимента с наибольшей вероятностью, а модой непрерывной случайной величины – значение, соответствующее максимуму плотности вероятности $f(x)$.

***Медианой*** непрерывной случайной величины $X$ называется такое ее значение $Me$, для которого вероятности того, что $X < Me$ и $X > Me$, равны:

$$P(X < Me) = P(X > Me) = 0{,}5.$$

Из определения функции распределения $F(X)$ следует, что $F(Me) = 0{,}5$. В случае симметричного распределения медиана совпадает с модой и математическим ожиданием.

***Квантилем*** $x_p$ является такое значение аргумента функции распределения, при котором выполняется равенство

$$F(x_p) = p.$$

Очевидно, что квантиль $x_{0,5}$ – это медиана распределения, т. к. $F(x_{0,5}) = 0{,}5$. Квантиль $x_{0,25}$ называется первым (или нижним) квантилем, а квантиль $x_{0,75}$ – третьим (или верхним) квантилем. Из выражения (2.1) следует, что

$$P(x_p < X < x_q) = q - p.$$

***Дисперсия*** $D(X)$ дискретной случайной переменной $X$ равна:

$$D(X) = \sum_{i=1}^{n} (x_i - m_x)^2 \, p_i.$$

Для непрерывной случайной величины дисперсия $D(X)$ определяется следующим выражением:

$$D(X) = \int_{-\infty}^{\infty} (x - m_x)^2 f(x) dx = M\{(x - m_x)^2\}.$$

Очевидно, что дисперсия – это центральный момент 2-го порядка, равный математическому ожиданию квадрата отклонения случайной величины $X$ от ее математического ожидания.

Дисперсию случайной величины $D(X)$ удобно вычислять по следующим формулам:

1) для дискретной величины

$$D(X) = M(x^2) - {m_x}^2 - \sum_{i=1}^{n} {x_i}^2 p_i - {m_x}^2;$$

2) для непрерывной случайной величины

$$D(X) = \int_{-\infty}^{\infty} x^2 f(x) dx - {m_x}^2.$$

Дисперсия обладает следующими свойствами:

1) $D(C) = 0$, где $C = \text{const}$;
2) $D(C \cdot X) = C^2 \cdot D(X)$;
3) для независимых случайных величин $X$ и $Y$ справедливо равенство

$$D(X \pm Y) = D(X) \pm D(Y).$$

***Средним квадратическим отклонением*** $\sigma(X)$ (среднеквадратическим отклонением или стандартом случайной величины $X$) называется арифметический корень из дисперсии:

$$\sigma(X) = \sigma_x = \sqrt{D(X)}\ .$$

Удобство использования величины $\sigma_x$ по сравнению с дисперсией $D(X)$ для оценки рассеивания случайной величины $X$ обусловлено тем, что размерность $\sigma_x$ совпадает с размерностью самой случайной величины $X$.

***Коэффициент вариации*** $V(X)$, характеризующий уровень рассеивания (вариацию) случайной переменной относительной ее математического ожидания, равен:

$$V(X) = \left|\frac{\sigma_x}{m_x}\right| \cdot 100\%.$$

***Коэффициент асимметрии*** $As$ задается формулой

$$As = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\sqrt{D^3}},$$

где $\mu_3$ – центральный момент 3-го порядка.

Значение коэффициента асимметрии $As$ характеризует степень асимметрии распределения относительно математического ожидания. Для любого симметричного распределения $As = 0$. Если коэффициент асимметрии отрицателен ($As < 0$), то либо большая часть значений дискретной случайной величины, либо ее мода находятся левее математического ожи-

дания. Справедливо и обратное утверждение: если коэффициент асимметрии положителен ($As > 0$), то либо большая часть значений дискретной случайной величины, либо ее мода находятся правее математического ожидания.

В случае непрерывной асимметрично распределенной случайной величины при $As > 0$ более пологая часть графика плотности распределения располагается правее моды, т. е. имеет место правосторонняя асимметрия, а при $As < 0$ – левее моды, т. е. имеет место левосторонняя асимметрия (рис. 2.2.).

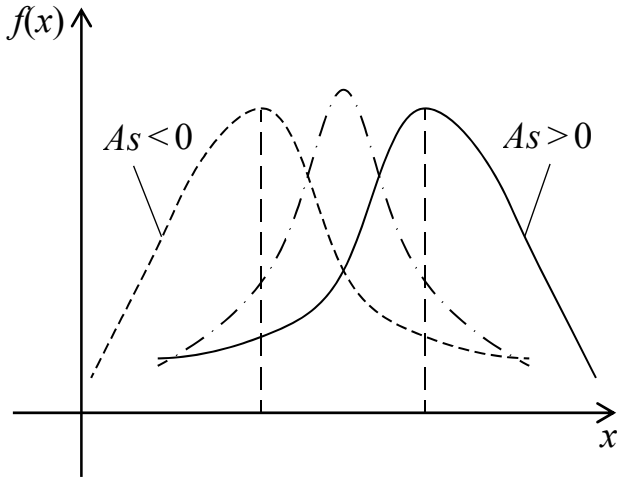


Рис. 2.2. Графики плотности распределения
непрерывной случайной переменной при $As > 0$ и $As < 0$

***Коэффициент эксцесса*** $Ex$ рассчитывается по формуле

$$Ex = \frac{\mu_4}{D^2} - 3 = \frac{\mu_4}{\sigma^4} - 3,$$

где $\mu_4$ – центральный момент 4-го порядка.

В случае непрерывной случайной величины, для которой $Ex > 0$, график плотности распределения имеет более острую вершину, чем для случайной величины, распределенной по нормальному закону (распределение Гаусса, для которого $Ex = 0$). Распределения с $Ex < 0$ более плосковершинные (рис. 2.3.). В случае дискретной случайной величины с $Ex > 0$ полученные в ходе эксперимента данные сконцентрированы около значения математического ожидания, а при $Ex < 0$ – более равномерно распределены по всей области их возможных значений.

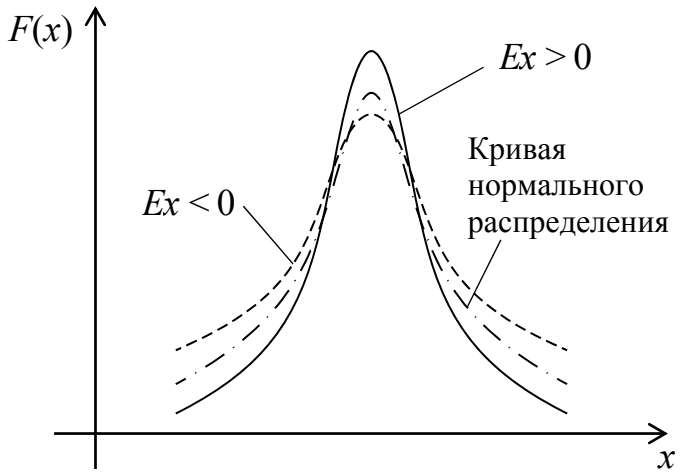


Рис. 2.3. Графики плотности распределения
непрерывной случайной переменной при $Ex > 0$ и $Ex < 0$

Обработка результатов эксперимента может быть в значительной мере упрощена, если от переменной $x$, записанной в натуральном масштабе, перейти к нормированной переменной $x_0$, используя следующее преобразование:

$$x_0 = \frac{x - m_x}{\sigma_x}.$$

Учитывая рассмотренные выше свойства математического ожидания и дисперсии, можно установить, что

$$M\{x_0\} = M\left\{\frac{x - m_x}{\sigma_x}\right\} = \frac{1}{\sigma_x} M\{x - m_x\} = \frac{1}{\sigma_x}[M\{x\} - M\{m_x\}] =$$

$$= \frac{1}{\sigma_x}(m_x - m_x) = 0,$$

$$D\{x_0\} = D\left\{\frac{x - m_x}{\sigma_x}\right\} = \frac{1}{\sigma_x^2} D\{x - m_x\} = \frac{1}{\sigma_x^2} D[D(x) - D(x)] =$$

$$= \frac{1}{\sigma_x^2}(\sigma_x^2 - 0) = 1.$$

## 2.4. Нормальное распределение случайной переменной

Нормальное распределение, называемое также распределением Гаусса, имеет большое значение в теории планирования эксперимента, проводимого для построения стохастической модели ХТО. Согласно центральным предельным теоремам теории вероятности, сумма большого количества

произвольным образом распределенных независимых случайных величин, среди которых нет доминирующей, имеет распределение, близкое к нормальному. Это в полной мере соответствует приведенному выше описанию обобщенной структуры ХТО (см. рис. 1.1.), подверженного влиянию большого числа неконтролируемых случайных воздействий.

Плотность нормального распределения (распределения Гаусса) определяется выражением

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}.$$

Данное распределение относится к классу двухпараметрических, т. к. оно полностью задается двумя параметрами: математическим ожиданием $m_x$ и дисперсией $\sigma_x^2$.

Нормальное распределение симметрично относительно математического ожидания, следовательно, коэффициент асимметрии этого распределения равен нулю ($As = 0$), равен нулю и коэффициент эксцесса ($Ex = 0$), а значения моды *Mo*, медианы *Me* и математического ожидания $m_x$ совпадают.

Нормальное распределение теоретически разработано наиболее полно, кроме того, с ним связаны несколько других распределений, используемых при обработке результатов эксперимента.

В случае нормального распределения нормированной случайной величины $x_0$ функция плотности распределения определяется выражением

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

а функция распределения, согласно формуле (2.3), определяется выражением

$$F_0(x) = \int_{-\infty}^{x} f_0(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} dx =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{x^2}{2}} dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{x^2}{2}} dx,$$

в котором первое слагаемое равно 0,5. Второе слагаемое в данном выражении называется ***функцией Лапласа*** $\Phi(x)$ или интегралом вероятности и определяется по формуле

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{x^2}{2}} dx.$$

При этом функция распределения нормированной нормально распределенной случайной величины $x_0$ равна:

$$F_0(x) = 0{,}5 + \Phi(x) = 0{,}5 + \Phi\left(\frac{x - m_x}{\sigma_x}\right).$$

Вероятность попадания случайной величины *X*, подчиняющейся нормальному закону, в интервал (*a*; *b*), равна:

$$P\{a \le x \le b\} = P\left\{\frac{a - m_x}{\sigma_x} \le x_0 \le \frac{b - m_x}{\sigma_x}\right\} = F_0(b) - F_0(a) =$$

$$= \Phi\left\{\frac{b - m_x}{\sigma_x}\right\} + 0{,}5 - \left[\Phi\left\{\frac{a - m_x}{\sigma_x}\right\} + 0{,}5\right] =$$

$$= \Phi\left\{\frac{b - m_x}{\sigma_x}\right\} - \Phi\left\{\frac{a - m_x}{\sigma_x}\right\}. \qquad (2.4.)$$

Воспользовавшись уравнением (2.4), можно определить вероятность события, заключающегося в том, что отклонение нормально распределенной случайной величины от ее математического ожидания не превысит величины $3\sigma_x$, т. е. вероятность равна:

$$P\{m_x - 3\sigma_x \le x \le m_x + 3\sigma_x\}.$$

Полагая в уравнении (2.4) $a = m_x - 3\sigma_x$, $b = m_x - 3\sigma_x$ и учитывая нечетность функции Лапласа, получаем:

$$P\{m_x - 3\sigma_x \le x \le m_x + 3\sigma_x\} = \Phi\{3\} - \Phi\{-3\} = 2\Phi\{3\} = 0{,}9973.$$

Событие с таким значением вероятности принято считать практически достоверным. Следовательно, можно сформулировать ***правило трех сигм***: вероятность того, что случайная величина отклонится от своего математического ожидания на величину большую, чем утроенное среднее квадратичное отклонение, практически равна нулю.

# 3. ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА

## 3.1. Классификация ошибок эксперимента на химико-технологическом объекте

В результате каждого конкретного эксперимента по определению значения переменной получают значение переменной $x$, отличающееся от ее истинного значения $x_{\text{ист}}$ на величину ошибки $\Delta x$, равную:

$$\Delta x = x_{\text{ист}} - x.$$

В зависимости от того, что является основной причиной появления ошибки, различают ***грубые***, ***систематические*** и ***случайные ошибки***.

Грубые ошибки (аномальные или резко выделяющиеся наблюдения) возникают вследствие нарушения условий проведения эксперимента, искажения его результатов в процессе контроля или регистрации оператором. С целью выявления и исключения таких результатов разработан ряд критериев для резко выделяющихся наблюдений.

Систематические ошибки присутствуют во всех результатах в виде постоянной составляющей или величины, изменяющейся по определенному закону. Их наличие обычно легко устанавливается и устраняется введением случайных поправок в результаты эксперимента.

Случайные ошибки сохраняются в результатах после отсеивания грубых и систематических ошибок. Их появление связанно с воздействием на ХТО большого количества случайных факторов, которые по разным причинам не могут быть выявлены и отнесены к входным воздействиям объекта. Полагается, что случайные ошибки распределены симметрично относительно значения $x_{\text{ист}}$, т. е. $M\{\Delta x\} = 0$.

## 3.2. Определение выборочных параметров распределения

При статистическом анализе необходимо четко разграничивать ***выборочные параметры*** и ***параметры генеральной совокупности***. Под генерированной совокупностью понимается гипотетическая область, состоящая из всех возможных значений переменных ХТО. Вследствие влияния на объект неконтролируемых, случайным образом изменяющихся воздействий $\gamma_i (i = 1, \dots, r)$ его переменные, к которым относятся как факторы $x_i$ ($i$=1, …, $k$), так и функции отклика $y_i$ ($i$=1, …, $m$), принимают некоторые случайные значения $x_{ij}$ и $y_{ij}$, которые называются реализациями случайных переменных $x_i$ и $y_i$. В результате проведения $n$ независимых опытов формируются случайные выработки из генеральной совокупности (выборочные совокупности): $x_{i1}, x_{i2}, \dots , x_{in}$ и $y_{i1}, y_{i2}, \dots , y_{in}$.

Число проведенных экспериментов называется ***объемом выборки***.

Задача статистического анализа состоит в том, чтобы оценить параметры генерированной совокупности по полученной выработке. При этом очевидно, что любая такая оценка имеет вероятностный характер и при обработке результатов эксперимента необходимо стремиться к обеспечению максимально возможной точности статистических оценок. Для этого выработка должна быть ***представительной*** (репрезентативной), т. е. должна отражать все закономерности, присущие генеральной совокупности. Очевидно, что увеличение объема выборки при прочих равных условиях повышает точность оценивания, но такой подход зачастую сопряжен с недопустимыми затратами средств и времени на проведение эксперимента.

Поэтому при обработке результатов эксперимента необходимо использовать методы, которые либо обеспечивают повышение точности оценок при фиксированном объеме выборки, либо допускают его уменьшение при сохранении заданной точности.

К числу условий формирования репрезентативной выборки можно отнести использование различных процедур ***рандомизации*** при проведении эксперимента, а также методов, позволяющих исключить из выборки грубые и систематические ошибки.

Рандомизацией называется любая процедура, обеспечивающая случайный порядок проведения опытов, включенных в план эксперимента на ХТО.

### 3.3. Точечные и интервальные оценки

Пусть $a$ – неизвестный параметр генеральной совокупности (например, математическое ожидание $m_x$ или дисперсия $\sigma_x^2$). По выборке из генеральной совокупности может быть вычислена только его оценка $\tilde{a}$, являющаяся случайной величиной, т. к. сама выборка случайна. Закон распределения $f(x, a)$ в общем случае зависит от закона распределения переменной $f(x)$, неизвестного генерального параметра $a$ и объема выборки $n$. Если эта оценка определяется одним числом, то ее называют ***точечной оценкой***. Так, оценкой математического ожидания случайной величины $x$, распределенной по нормальному закону, является среднее арифметическое значение результатов эксперимента $x_i$ ($i = \overline{1, n}$), представленных в выборке:

$$m_x \to \bar{x} = \frac{1}{n}\sum_{i=1}^{n}. \tag{3.1}$$

В связи с этим математическое ожидание случайной величины $m_x$ иногда называют ***генеральным средним***.

Оценкой генеральной дисперсии $\sigma_x^2$ является выборочная дисперсия $s_x^2$:

$$\sigma_x^2 \to s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.2}$$

В общем случае выборочная дисперсия представляет собой отношение суммы квадратов отклонений результатов эксперимента $ss$ от некоторого уровня к числу степеней свободы $f$:

$$s^2 = \frac{ss}{f}. \tag{3.3}$$

Число степеней свободы определяется как разность между объемом выборки $n$ и ***числом наложенных связей***, равным количеству статистик,

рассчитанных по той же выборке, по которой рассчитывается дисперсия. В выражении (3.3) сумма квадратов отклонений равна:

$$ss = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

при этом число степеней свободы $f = n - 1$, т. к. при вычислении дисперсии единственная наложенная связь – среднее значение $\bar{x}$, рассчитываемое по той же выборке, что и $s_x^2$.

Оценка для среднеквадратического отклонения (выборочный стандарт распределения) равна:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

К точечной оценке $\tilde{a}$ генерального параметра $a$ предъявляется ряд требований, при выполнении которых можно утверждать, что оценка обеспечивает достаточно хорошие приближения оцениваемых параметров:

1) При увеличении объема выборки оценка $\tilde{a}$ должна приближаться (сходится по вероятности) к оцениваемому параметру $a$. Такая оценка называется ***состоятельной***.

2) Желательно также, чтобы не происходило систематического занижения или завышения оценки, рассчитанной по выборкам из генеральной совокупности, по сравнению с истинным значением генерального параметра, т. е. необходимо, чтобы $M\{\tilde{a}\} = a$. При выполнении этого условия оценка является несмещенной. Оценка математического ожидания, определяемого по формуле (3.1), является несмещенной и состоятельной.

Приведенная оценка дисперсии, определяемая по формуле (3.2), также является несмещенной.

3) Если по заданной выборке могут быть вычислены несколько различных оценок $\tilde{a}$, то та из них, для которой дисперсия $D\{\tilde{a}\}$ минимальна, называется ***эффективной*** и является наиболее предпочтительной. Например, оценкой математического ожидания наряду с оценкой, определяемой по формуле (3.1), может служить среднее арифметическое наибольшего и наименьшего значений в выборке ($x_{\max}$ и $x_{\min}$ соответственно), но дисперсия такой оценки будет выше по сравнению с дисперсией оценки, полученной из формулы (3.1):

$$\tilde{x} = \frac{x_{\min} + x_{\max}}{2}.$$

Для выборок небольшого объема точечные оценки ненадежны. Если такая выборка содержит аномальные наблюдения, то расхождение между генеральным параметром и его оценкой будет значительным. Поэтому для таких выборок предпочтительным является ***интервальное оценивание***.

***Доверительным интервалом*** для генерального параметра $a$ называют такой интервал $(\tilde{a}_1; \tilde{a}_2)$, которому с заранее выбранной вероятностью $p$, близкой к единице, принадлежит неизвестное значение параметра $a$:

$$P\{\tilde{a}_1 < a < \tilde{a}_2\} = p,$$

где $p$ – ***доверительная вероятность***.

Выбор величины доверительной вероятности определяется экспериментатором в зависимости от специфики объекта исследования. Обычно принимают $p$ = 0,95; 0,99; 0,90. Чем больше доверительная вероятность при неизменном объеме эксперимента, тем больше интервал. При увеличении числа опытов и фиксированном значении $p$ ширина доверительного интервала уменьшается.

Вероятность события, противоположного рассмотренному событию (когда генеральный параметр $a$ не попадает в доверительный интервал), определяется из выражения

$$\alpha = 1 - p,$$

где $\alpha$ – уровень значимости.

В качестве границ доверительного интервала обычно берут симметричные квантили, т. е. $\tilde{a}_1 = \tilde{a}_{\alpha/2}$ и $\tilde{a}_2 = \tilde{a}_{1-\alpha/2}$. Значения указанных квантилей не могут быть найдены по плотности распределения $f(x, a)$ выборочной оценки $\tilde{a}$, т. к. ее закон распределения зависит не только от закона распределения случайных значений результатов эксперимента, но и от неизвестного значения генерального параметра $a$. Для выборок достаточно большого объема ($n \geq 50$) в качестве значения $a$ можно использовать его точечную оценку $\tilde{a}$.

В случае, когда количество наблюдений небольшое, для нахождения границ доверительного интервала рассматривают не распределение оценки $\tilde{a}$, а распределение вспомогательной величины $\theta$, которое зависит только от распределения результатов эксперимента и от объема выборки. Задавшись значением доверительной вероятности (или уровнем значимости), предварительно определяют границы интервала для величины $\theta$, которые затем пересчитываются в границы искомого интервала $(\tilde{a}_1; \tilde{a}_2)$.

***Пример 1:*** При проведении процесса щелочной делигнификации целлюлозы были получены следующие данные о плотности черного щелока: $X = \{x_i\} = \{1{,}05;\ 1{,}07;\ 1{,}11;\ 1{,}09\}$ г/м$^3$.

Необходимо с доверительной вероятностью $p$ =0,95 произвести интервальное оценивание математического ожидания $m_x$ указанного технологического параметра.

Оцениваемый генеральный параметр присутствует в выражении для вспомогательной величины $t$, которая в случае нормального распределения выборки *X* имеет распределение Стьюдента (*t*-распределение):

$$t = \frac{\bar{x} - m_x}{s_x}\sqrt{n},$$

Плотность *t*-распределения зависит только от числа свободы $f$:

$$f = n - 1,$$

где $n$ – объем выборки.

Доверительный интервал для переменной $t$ равен:

$$t_{f,\alpha/2} \le \frac{\bar{x} - m_x}{s_x}\sqrt{n} \le t_{f,1-\alpha/2}, \qquad (3.4)$$

где $t_{f,\alpha/2}$, $t_{f,1-\alpha/2}$ – квантили *t*-распределения.

В силу симметрии этого распределения ($t_{f,\alpha/2} = -t_{f,1-\alpha/2}$) неравенство (3.4) можно записать в следующем виде:

$$-t_{f,1-\alpha/2} \le \frac{\bar{x} - m_x}{s_x}\sqrt{n} \le t_{f,1-\alpha/2},$$

а затем преобразовать его в неравенство, определяющее доверительный интервал для определяемого генерального параметра $m_x$:

$$\bar{x} - \frac{s_x}{\sqrt{n}} t_{f,1-\alpha/2} \le m_x \le \bar{x} + \frac{s_x}{\sqrt{n}} t_{f,1-\alpha/2}. \qquad (3.5)$$

Среднее значение, определенное по результатам четырех измерений, равно: $\bar{x}$ =1,08 г/м$^3$.

Выборочный стандарт распределения равен: $s_x$ = 0,026 г/м$^3$.

Квантиль *t*-распределения, соответствующий числу степеней свободы $f = 3$ и уровню значимости α =0,05, равен: $t_{3;0,975}$=3,182 [7].

Тогда, согласно формуле (3.5), получаем доверительный интервал для генерального среднего:

$$1{,}039 \le m_x \le 1{,}121.$$

***Пример 2:*** При определении фосфора в образцах твердого топлива фотоколориметрическим методом по ГОСТ 1932-93 «Топливо твердое. Методы определения фосфора» были получены результаты, приведенные в табл. 3.1.

Таблица 3.1

Данные по определению интервальной оценки генеральной дисперсии

| Номер опыта | Содержание фосфора, % масс. | Номер опыта | Содержание фосфора, % масс. | Номер опыта | Содержание фосфора, % масс. | Номер опыта | Содержание фосфора, % масс. |
|---|---|---|---|---|---|---|---|
| 1 | 0,015 | 9 | 0,023 | 17 | 0,016 | 25 | 0,014 |
| 2 | 0,017 | 10 | 0,017 | 18 | 0,015 | 26 | 0,016 |
| 3 | 0,016 | 11 | 0,019 | 19 | 0,02 | 27 | 0,017 |
| 4 | 0,012 | 12 | 0,015 | 20 | 0,019 | 28 | 0,019 |
| 5 | 0,015 | 13 | 0,014 | 21 | 0,018 | 29 | 0,015 |
| 6 | 0,014 | 14 | 0,022 | 22 | 0,016 | 30 | 0,013 |
| 7 | 0,016 | 15 | 0,013 | 23 | 0,015 | 31 | 0,016 |
| 8 | 0,019 | 16 | 0,018 | 24 | 0,021 | 32 | 0,017 |

Необходимо найти доверительный интервал для генеральной дисперсии при доверительной вероятности 90 %.

Если выборка $X = \{x_i\}$ сформирована из нормально распределенной генеральной совокупности с дисперсией $\sigma_x^2$, то сумма $\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)^2$ распределена по закону $\chi^2$ с числом степеней свободы $f = n - 1$, причем плотность распределения $\chi^2$ зависит только от числа степеней свободы $f$ [2].

Преобразовав выражение для $\chi^2$ с учетом формулы (3.2), получим:

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)^2 = \frac{(n-1)\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma_x^2(n-1)} = \frac{f}{\sigma_x^2}\frac{\sum(x_i-\bar{x})^2}{n-1} = \frac{fs_x^2}{\sigma_x^2}.$$

Задаваясь величиной уровня значимости $\alpha$, находим границы доверительного интервала для переменной $\chi^2$:

$$\chi^2_{f,\alpha/2} \le \chi^2 \le \chi^2_{f,1-\alpha/2}$$

или

$$\chi^2_{f,\alpha/2} \le \frac{fs_x^2}{\sigma_x^2} \le \chi^2_{f,1-\alpha/2}. \qquad (3.6)$$

Преобразуя неравенство (3.6), получаем выражение, определяющее доверительный интервал для генеральной дисперсии $\sigma_x^2$:

$$\frac{fs_x^2}{\chi^2_{f,1-\alpha/2}} \le \sigma^2 \le \frac{fs_x^2}{\chi^2_{f,\alpha/2}}.$$

Для рассматриваемого примера $\bar{x} = 0{,}017;\ s_x^2 = 6{,}9 \cdot 10^{-6}$.

Для заданной доверительной вероятности $p = 0{,}9$ и числа степеней свободы $f = 31$ находим квантили распределения $\chi^2$ [7]: $\chi^2_{31;0{,}025} = 17{,}54$; $\chi^2_{31;0{,}975} = 44{,}99$.

В результате имеем:

$$4{,}75 \cdot 10^{-6} \leq \sigma^2 \leq 1{,}22 \cdot 10^{-5}.$$

Доверительный интервал для среднеквадратического отклонения равен:

$$2{,}18 \cdot 10^{-3} \leq \sigma \leq 3{,}49 \cdot 10^{-3}.$$

### 3.4. Предварительная обработка экспериментальных данных

Решению задачи построения статистической модели ХТО обязательно предшествует этап предварительной обработки экспериментальных данных.

Этот этап включает в себя следующие задачи:

1) отсеивание грубых погрешностей и устранение систематических ошибок, содержащихся в результатах эксперимента;

2) проверку соответствия результатов эксперимента нормальному распределению. В том случае, когда распределение полученных результатов отличается от нормального, следует определить, какому закону распределения они подчиняются, и, если это возможно, преобразовать данное распределение к нормальному.

Решение указанных задач этапа предварительной обработки экспериментальных данных основывается на проверке соответствующих статистических гипотез.

### 3.5. Проверка статистических гипотез

Статистическая гипотеза – это некоторое подлежащее проверке предположение (высказывание) о:

1) параметрах распределения генеральной совокупности (***параметрическая гипотеза***);

2) законе распределения генеральной совокупности (***гипотеза согласия***).

Уже на этапе предварительной обработки экспериментальных данных необходимо осуществлять проверку статистических гипотез о наличии в результатах эксперимента грубых ошибок, а после их отсеивания – проводить проверку гипотезы о том, что результаты эксперимента распределены по нормальному закону.

При проверке гипотезы, как и при интервальном оценивании, необходимо предварительно задаться доверительной вероятностью $p$, которой соответствует уровень значимости $\alpha$: $\alpha = 1 - p$. Для выдвинутой гипотезы $H_0$ (***нулевой гипотезы***) необходимо выбрать соответствующую статистику $\tilde{\theta}$, называемую ***критерием проверки*** или ***критерием значимости***. Критерием проверки называют также и методику проверки гипотезы. На основании принятой гипотезы о распределении критерия проверки $\tilde{\theta}$, задавшись значениями уровня значимости $\alpha$ и числом степеней свободы $f$, находят границы доверительного интервала $(\tilde{\theta}_1;\ \tilde{\theta}_2)$, который называют ***областью принятия гипотезы***.

Далее необходимо вычислить значение критерия проверки $\theta$ по данным выборки. В зависимости от величины $\theta$ по отношению к области принятия гипотезы нулевая гипотеза или принимается, или отвергается. Если $\tilde{\theta}_1 < \theta < \tilde{\theta}_2$, то на принятом уровне значимости нет основания для отвержения гипотезы и она принимается. Если вычисленное по выборке значение критерия $\theta < \tilde{\theta}_1$ или $\theta > \tilde{\theta}_2$, т. е. значение $\theta$ попадает в ***критическую область***, то гипотеза отвергается. В последнем случае принимается одна или несколько ***альтернативных (конкурирующих) гипотез*** $H_1, H_2, \ldots$ .

Выбор альтернативной гипотезы зависит от конкретной формулировки задачи. Например, если $H_0: \sigma_1^2 = \sigma_2^2$, т. е. дисперсии двух генеральных совокупностей равны, то $H_1: \sigma_1^2 \neq \sigma_2^2$. Той же нулевой гипотезе $H_0$ могут соответствовать две альтернативные гипотезы: $H_1: \sigma_1^2 < \sigma_2^2$ и $H_2: \sigma_1^2 > \sigma_2^2$. В этом случае гипотеза называется ***двусторонней***. Если выполнение одного из этих двух неравенств по каким-либо причинам невозможно, то альтернативная гипотеза является ***односторонней***.

Поскольку значение критерия проверки гипотезы $\theta$ рассчитывается по выборочной совокупности, которая является случайной, то, принимая или отклоняя выдвинутую гипотезу $H_0$, можно допустить ошибки двух типов. ***Ошибка первого рода*** состоит в том, что нулевая гипотеза отвергается и принимается гипотеза $H_1$, в то время как в действительности верна гипотеза $H_0$. ***Ошибка второго рода*** заключается в том, что принимается гипотеза $H_0$, в то время как верна гипотеза $H_1$.

Вероятность ошибки первого рода равна уровню значимости $\alpha$. Вероятность ошибки второго рода $\beta$ зависит от многих факторов. Величину $(1 - \beta)$ называют ***мощностью критерия***. Если выдвинутую гипотезу можно проверить с помощью различных критериев, то выбирают тот из них, которому при заданном уровне значимости $\alpha$ соответствует большая мощность.

Таким образом, процедура проверки параметрической статистической гипотезы предусматривает последовательное выполнение следующих этапов:

1) формулируется гипотеза $H_0$, а также подлежащая проверке и конкурирующая гипотеза $H_1$;

2) выбирается значение доверительной вероятности $p$ или уровня значимости α;

3) устанавливается распределение статистики, соответствующей используемому критерию проверки при условии, что проверяемая гипотеза $H_0$ верна;

4) исходя из закона распределения выбранной статистики и из того, как сформулирована альтернативная гипотеза, определяется критическая область;

5) по данным выборки рассчитывается значение статистики θ;

6) формулируется заключение по проверке гипотезы: если рассчитанное значение статистики попадает в критическую область, то гипотеза $H_0$ отклоняется, в противном случае гипотеза принимается.

### *3.5.1. Проверка гипотезы об однородности двух выборочных дисперсий*

Пусть имеется две случайные выборки объемами $n_1$ и $n_2$. Для них рассчитаны средние значения $\bar{x}_1, \bar{x}_2$ и выборочные дисперсии $s_1^2, s_2^2$. Необходимо проверить гипотезу о равенстве генеральных дисперсий $\sigma_1^2$ и $\sigma_2^2$, для которых $s_1^2$ и $s_2^2$ являются точечными оценками. Следовательно, проверяемая гипотеза $H_0: \sigma_1^2 = \sigma_2^2$.

Для проверки выдвинутой гипотезы следует воспользоваться распределением Фишера (*F*-распределением) для случайной величины:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}. \tag{3.7}$$

Вид этого распределения зависит только от числа степеней свободы $f_1 = n_1 - 1$ и $f_2 = n_2 - 1$.

Поскольку проверяется гипотеза $\sigma_1^2 = \sigma_2^2$, то ***F-отношение*** (3.7) принимает следующий вид:

$$F = \frac{s_1^2}{s_2^2}. \tag{3.8}$$

Область принятия гипотезы для выбранного уровня значимости α определяется неравенством

$$F_{\frac{\alpha}{2},f_1,f_2} \leq \frac{S_1^2}{S_2^2} \leq F_{1-\frac{\alpha}{2},f_1,f_2}. \tag{3.9}$$

Особенностью *F*-распределения является то, что

$$F_{\frac{\alpha}{2},f_1,f_2} = \frac{1}{F_{1-\frac{\alpha}{2},f_1,f_2}},$$

поэтому неравенство (3.8) можно преобразовать к виду

$$\frac{1}{F_{1-\frac{\alpha}{2},f_1,f_2}} \leq \frac{s_1^2}{s_2^2} \leq F_{1-\frac{\alpha}{2},f_1,f_2}. \quad (3.10)$$

В случае проверки односторонней гипотезы областью ее принятия будет служить одно из следующих неравенств:

$$\frac{S_1^2}{S_2^2} \geq \frac{1}{F_{1-\alpha,f_1,f_2}}, \qquad \frac{S_1^2}{S_2^2} \leq F_{1-\alpha,f_1,f_2}. \quad (3.11)$$

Если рассчитанная по выборкам величина *F*-отношения [см. уравнение (3.8)] удовлетворяет условиям (3.10) или (3.11), то гипотеза $H_0$ принимается, а различие между точечными оценками $s_1^2$ и $s_2^2$ признается случайным. В этом случае выборочные дисперсии называются ***однородными***. При попадании *F*-отношения в критическую область гипотеза $H_0$ на уровне значимости α отклоняется и различие между дисперсией $s_1^2$ и $s_2^2$ следует признать статистически значимым.

При использовании данного критерия в качестве $s_1^2$ обычно рассматривают большую из двух выборочных дисперсий.

***Пример 3:*** При определении содержания марганца в питьевой воде с помощью атомно-абсорбционной спектрофотомерии (метод I) и атомно-эмиссионной спектрометрии (метод II) были получены следующие результаты:

1) по методу I: 0,048; 0,049; 0,043; 0,047; 0,044 мг/дм$^3$;
2) по методу II: 0,044; 0,045; 0,042; 0,043; 0,044; 0,042 мг/дм$^3$.

Необходимо сравнить точности двух методов измерения при доверительной вероятности 95 %.

Задача, поставленная в данном примере, предполагает необходимость проверки гипотезы о равенстве генеральных дисперсий представленных выборок, т. е. $H_0: \sigma_1^2 = \sigma_2^2$ на уровне значимости $\alpha = 0{,}05$.

Т. к. в приведенной постановке задачи нет указания на то, что один из методов априори является более точным, то данная гипотеза является двусторонней и область принятия гипотезы имеет вид выражения (3.10). В случае, когда по условию задачи необходимо было бы по данным выборки подтвердить одностороннюю гипотезу о том, что один из методов является более точным, область принятия гипотезы имела бы вид выражения (3.11).

Границы доверительного интервала, согласно формуле (3.10), для $\alpha = 0{,}05,\ f_I = 4$ и $f_{II} = 2$ равны:

$$F_{0{,}975;4;5} = 7{,}39,$$

$$\frac{1}{F_{0{,}975;4;5}} = 0{,}135.$$

Следовательно, имеем область принятия гипотезы:

$$0{,}139 \le \frac{s_I^2}{s_{II}^2} \le 7{,}39. \qquad (3.12)$$

Значение критерия проверки (*F*-отношение), соответствующее рассчитанным по выборкам дисперсиям воспроизводимости указанных методов $s_I^2 = 6{,}70 \cdot 10^{-6}$ и $s_{II}^2 = 1{,}47 \cdot 10^{-6}$, равно:

$$F = \frac{s_I^2}{s_{II}^2} = \frac{6{,}70 \cdot 10^{-6}}{1{,}47 \cdot 10^{-6}} = 4{,}57.$$

Поскольку *F*-отношение удовлетворяет неравенству (3.12), гипотеза о равенстве генеральных дисперсий принимается, выборочные дисперсии $s_I^2$ и $s_{II}^2$ полагаются однородными, а рассматриваемые в примере методы определения содержания марганца в питьевой воде – равноточными.

### *3.5.2. Проверка гипотезы об однородности n выборочных дисперсий (n>2)*

При обработке результатов эксперимента необходимость в проверке указанной гипотезы возникает, когда на ХТО проводится несколько ($n$) серий параллельных опытов с неизменным сочетанием значений факторов в каждой серии (осуществляется дублирование опытов). При этом общая дисперсия эксперимента – ***дисперсия воспроизводимости***, вычисленная по выражению

$$s^2_{\text{воспр}} = \frac{f_1 s_1^2 + f_2 s_2^2 + \cdots + f_k s_n^2}{f_1 + f_2 + \cdots f_n} = \frac{\sum_{i=1}^{n} f_i s_i}{f_{\text{воспр}}}, \qquad (3.13)$$

отражает воспроизводимость результатов эксперимента на объекте, но только при условии, что выборочные дисперсии серий параллельных опытов $s_i^2$ ($i$=1, …, $n$) однородны, т. е. их различие статистически не значимо.

Число степеней свободы дисперсии воспроизводимости, определяемой по формуле (3.13), равно:

$$f_{\text{воспр}} = \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} (n_i - 1) = \sum_{i=1}^{n} n_i - n = N - n, \qquad (3.14)$$

где $n_i$ – объем -й серии параллельных опытов; $N$ – общее по всем сериям число опытов в эксперименте на объекте.

Проверка гипотезы однородности $n$ выборочных дисперсий может быть проведена с использованием критерия Бартлета. Критерием проверки является отношение $B / C$, представляющее собой случайную величину, распределенную по закону $\chi^2$ с числом степеней свободы $f = n - 1$.

Составляющие критерии:

$$\left.\begin{aligned} B &= 2{,}303\left(f_{\text{воспр}} \log s^2_{\text{воспр}} - \sum_{i=1}^{n} f_i \log s_i^2\right), \\ C &= 1 + \frac{1}{3(n-1)}\left(\sum_{i=1}^{n} \frac{1}{f_i} - \frac{1}{f_{\text{воспр}}}\right). \end{aligned}\right\} \qquad (3.15)$$

Гипотеза об однородности дисперсий $s_i^2$ принимается, если для принятого уровня значимости α выполняется следующее неравенство:

$$B < \chi^2_{\alpha;n-1}. \qquad (3.16)$$

***Пример 4:*** Определение содержания нефтепродуктов в питьевой воде методом ИК-спектрометрии проводилось четырьмя лабораториями с получением следующих результатов:

1) лаборатория 1: 0,06; 0,07; 0,06; 0,05; 0,07;
2) лаборатория 2: 0,06; 0,08; 0,04; 0,09; 0,05; 0,06;
3) лаборатория 3: 0,05; 0,07; 0,05; 0,06;
4) лаборатория 4: 0,06; 0,07; 0,06; 0,10; 0,08.

Необходимо с доверительной вероятностью $p = 0{,}95$ установить, можно ли считать равными оценки случайных погрешностей результатов анализа разных лабораторий.

В соответствии с постановкой задачи необходимо проверить гипотезу об однородности четырех выборочных дисперсий по выборкам с объемами $n_1 = 5$, $n_2 = 6$, $n_3 = 4$ и $n_4 = 5$. Дисперсии, рассчитанные по результатам параллельных опытов, проведенных четырьмя лабораториями, равны: $s_1^2 = 7{,}00 \cdot 10^{-5}$, $s_2^2 = 3{,}47 \cdot 10^{-4}$, $s_3^2 = 9{,}17 \cdot 10^{-5}$ и $s_4^2 = 2{,}80 \cdot 10^{-4}$ соответственно.

Числа степеней свободы выборок: $f_1 = 4$, $f_2 = 5$, $f_3 = 3$ и $f_4 = 4$ .

Общая дисперсия по всем измерениям, согласно формуле (3.14), равна:

$$s^2_{\text{воспр}} = \frac{f_1 s_1^2 + f_2 s_2^2 + f_3 s_3^2 + f_4 s_4^2}{f_1 + f_2 + f_3 + f_4} =$$

$$\frac{4 \cdot 7{,}0 \cdot 10^{-5} + 5 \cdot 3{,}47 \cdot 10^{-4} + 3 \cdot 9{,}17 \cdot 10^{-5} + 4 \cdot 2{,}8 \cdot 10^{-4}}{4 + 5 + 3 + 4} = 2{,}13 \cdot 10^{-4}.$$

Число степеней свободы дисперсии воспроизводимости, определяемой по формуле (3.13), равно: $f_{\text{воспр}} = 16$.

Проверку выдвинутой гипотезы осуществляем по критерию Бартлета, для чего, согласно формулам (3.15), по данным выборок рассчитываем статистику: $B \,/\, C = 3{,}12$.

Критическое значение критерия на уровне значимости $\alpha = 0{,}05$ и при количестве проверяемых на однородность дисперсий $n = 4$ равно: $\chi^2_{0,05;3}$ =7,8.

Поскольку для рассчитанного по выборкам значения статистики $B \,/\, C$ неравенство (3.16) выполняется, выборочные дисперсии могут быть признаны однородными, а оценки случайных погрешностей результатов ИК-спектрометрии, проведенной четырьмя лабораториями, – равными.

Проверка гипотезы об однородности $n$ выборочных дисперсий может быть осуществлена по достаточно простому критерию Кохрена:

$$G = \frac{s^2_{\max}}{s_1^2 + s_1^2 + \cdots + s_n^2},$$

где $s^2_{\max}$ – выборочная дисперсия, имеющая наибольшее значение.

Ограничение на использование этого критерия по сравнению с критерием Бартлета обусловлено тем, что все серии параллельных опытов должны иметь одинаковый объем $n_0$. Следовательно, число степеней свободы для каждой серии равно: $f_0 = n_0 - 1$. Случайная величина $G$ имеет распределение, зависящее только от количества выборочных дисперсий $n$ и числа степеней свободы $f_0$.

Если на принятом уровне значимости α выполняется неравенство

$$G < G_{\alpha,n,f_0,}$$

то дисперсии $s_i^2$ являются однородными и по ним может быть рассчитана дисперсия воспроизводимости $s^2_{\text{воспр}}$ с числом степеней свободы $f_{\text{воспр}}$:

$$f_{\text{воспр}} = n f_0 = n(n_0 - 1).$$

### *3.5.3. Проверка гипотезы о случайном характере различия средних значений результатов эксперимента в двух выборках*

Для эксперимента, состоящего из двух серий опытов, может возникнуть необходимость в проверке гипотезы о случайном характере различия средних в сериях, т. е. выдвигается гипотеза $H_0: m_{x1} = m_{x2}$ о равенстве математических ожиданий двух генеральных нормально распределенных совокупностей. Гипотеза проверяется на основании данных двух независимых выборок объемами $n_1$ и $n_2$. По данным выборок рассчитываются средние значения $\bar{x}_1$ и $\bar{x}_2$, являющиеся точечными оценками сравниваемых математических ожиданий, и выборочные дисперсии $s_{x1}^2$ и $s_{x2}^2$. Предполагается, что гипотеза об однородности выборочных дисперсий принимается и может быть рассчитана дисперсия воспроизводимости:

$$s_{\text{воспр}}^2 = \frac{f_1 s_{x1}^2 + f_2 s_{x2}^2}{f_1 + f_2},$$

где $f_1 = n_1 - 1$; $f_2 = n_2 - 1$.

Проверка производится с помощью статистики $t$:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_{\text{воспр}}^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \tag{3.17}$$

подчиняющейся в случае справедливости проверяемой гипотезы распределению Стьюдента с числом степеней свободы $f = n_1 + n_2 - 2$.

Выдвинутая гипотеза принимается, если на принятом уровне значимости выполняется неравенство

$$|t| < t_{1-\frac{\alpha}{2};f}. \tag{3.18}$$

Если гипотеза об однородности выборочных дисперсий $s_{x1}^2$ и $s_{x2}^2$ отвергается, то вместо статистики, определяемой по выражению (3.17), необходимо воспользоваться статистикой

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1} s_{x1}^2 + \frac{1}{n_2} s_{x2}^2}}.$$

При этом условие принятия гипотезы о случайном характере различия средних [формула (3.18)] остается неизменным, но в качестве числа степеней свободы $f$ принимается наибольшее целое число, не превосходящее величины, равной

$$\frac{(n_1-1)(n_2-1)\left(\frac{s_{x1}^2}{n_1}+\frac{s_{x2}^2}{n_2}\right)^2}{(n_1-1)\left(\frac{s_{x1}^2}{n_1}\right)^2+(n_2-1)\left(\frac{s_{x2}^2}{n_2}\right)^2}.$$

Если выборок больше, чем две, то при сравнении нескольких средних можно также использовать -распределение, проводя сравнение средних попарно, или применить другие критерии, например, множественный ранговый критерий Дункана [1; 3].

***Пример 5:*** В рассмотренном ранее примере 3, связанном с определением содержания марганца в питьевой воде с помощью атомно-абсорбционной спектрофотомерии (метод I) и атомно-эмиссионной спектрометрии (метод II) было установлено, что гипотеза о равенстве генеральных дисперсий принимается: выборочные дисперсии $s_I^2 = 6{,}70 \cdot 10^{-6}$ и $s_{II}^2 = 1{,}47 \cdot 10^{-6}$ полагаются однородными.

Необходимо проверить гипотезу о случайном характере различия средних значений, соответствующих указанным методам.

По данным примера 3 определены значения выборочных средних: $\bar{x}_1 = 0{,}046$ и $\bar{x}_2 = 0{,}043$. Дисперсия воспроизводимости, рассчитанная по формуле (3.13), равна: $s^2_{\text{воспр}} = 3{,}79 \cdot 10^{-6}$, а соответствующее значение критерия $t$ равно: $t = 2{,}431$.

Критическое значение критерия на уровне значимости $\alpha = 0{,}05$ для числа степеней свободы $f = 9$ составляет $t_{0{,}975;\,9}$ =2,262.

Так как на принятом уровне значимости неравенство (3.18) выполняется, то гипотеза о случайном характере различия средних значений принимается.

## 3.6. Критерии исключения грубых ошибок

Разработано множество критериев, позволяющих отсеять резко выделяющиеся наблюдения. Область применения каждого из них зависит от полноты информации, характера распределения случайной величины, значениях параметров генеральной совокупности и объема выборки.

***Критерий 1:*** Для выборки небольшого объема ($n \leq 25$) может быть вычислена статистика τ по следующей формуле:

$$\tau = \frac{\left|x_{\max(\min)} - \bar{x}\right|}{s_x},$$

где $x_{\max(\min)}$ – наибольшее (наименьшее) значение в выборке; $\bar{x}$– выборочная оценка математического ожидания; $s_x$ – выборочная оценка среднего квадратического отклонения.

Затем, задавшись достаточно высоким уровнем доверительной вероятности $p$ или уровнем значимости $\alpha = 1 - p$, необходимо по таблице [7] определить критическое значение $\tau_{\alpha,n}^{\text{кр}}$ и сравнить его с полученным значением $\tau$. Если выполняется неравенство $\tau < \tau_{\alpha,n}^{\text{кр}}$, то наблюдение $x_{\max(\min)}$ оставляют в выборке.

Если указанное неравенство не выполняется, то резко выделяющееся наблюдение отсеивают и рассматривают следующее наибольшее (наименьшее) значение $x$, но предварительно необходимо пересчитать величины $\bar{x}$ и $s_x$ для нового объема выборки.

***Критерий 2:*** Проверка гипотезы о наличии грубой ошибки может быть произведена с помощью статистики $\tau$, определяемой из уравнения

$$\tau = \frac{x_{\max} - x'_{\max}}{d},$$

где $x_{\max}$ – наибольшее значение в выборке; $x'_{\max}$ – следующее после $x_{\max}$ наибольшее значение в выборке; $d$ – размах выборки.

Размах выборки $d$ вычисляется по формуле

$$d = x_{\max} - x_{\min}. \tag{3.19}$$

Затем, задаваясь уровнем значимости $\alpha$, определяют критическое значение $\tau_{\alpha,n}^{\text{кр}}$.

Как и при использовании предыдущего критерия, проверяемый результат эксперимента должен быть исключен, если выполняется неравенство $\tau > \tau_{\alpha,n}^{\text{кр}}$.

## 3.7. Проверка гипотезы о близости выборочного распределения результатов эксперимента к нормальному распределению (критерии согласия)

### *3.7.1. Критерий среднего абсолютного значения*

Для выборки объемом $n$, распределенной с доверительной вероятностью $p$=0,95 по закону, близкому к нормальному, справедливо следующее неравенство:

$$\left|\frac{\text{САО}}{s} - 0{,}7979\right| < \frac{0{,}4}{\sqrt{n}},$$

где САО – среднее абсолютное отклонение, определяемое по формуле

$$\text{САО} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}.$$

### *3.7.2. Критерий согласия Пирсона (критерий $\chi^2$)*

Некоторое представление о предлагаемом графике плотности распределения $f(x)$ генеральной совокупности можно получить по виду выборочного распределения $\hat{f}(x)$, называемого ***гистограммой распределения***. Для его построения весь интервал изменения случайной величины от $x_{\min}$ до $x_{\max}$ разбивается на $k$ равных интервалов, где величина $k$ вычисляется по следующей формуле с последующим округлением до ближайшего целого числа:

$$k = 1 + 3{,}21 \log n.$$

При этом длина каждого интервала $\Delta x$ равна:

$$\Delta x = \frac{d}{k},$$

где $d$ – размах выборки [определяется по формуле (3.19)].

Пусть $n_i$ – количество результатов случайной переменной, попавших в -й интервал ($i = \overline{1,k}$). Относительная частота попадания в этот интервал равна $n_i / n$. На каждом из интервалов ординаты точек гистограммы, график которой приведен на рис. 3.1., неизменны и равны $n_i / n\Delta x$ .

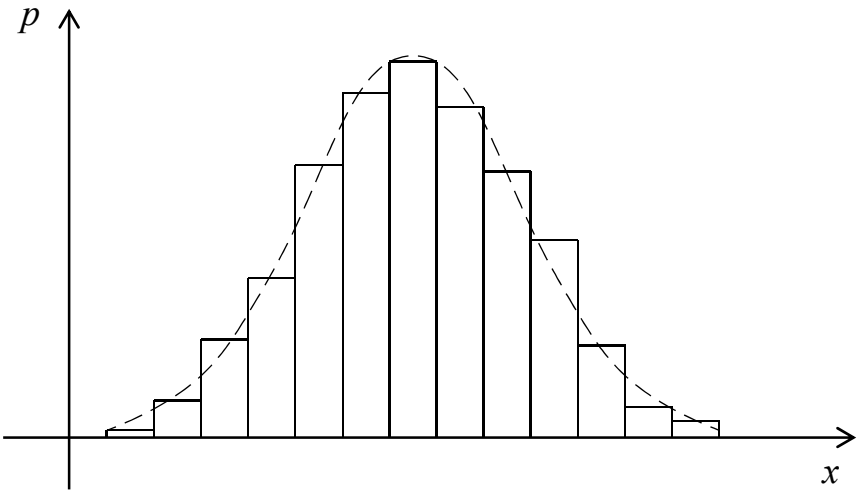


Рис. 3.1. Гистограмма выборочного распределения

Следуя проверяемой гипотезе о том, что функция плотности распределения генеральной совокупности равна $f(x)$, можно вычислить вероятность попадания случайной величины в каждый из интервалов. Так, для -го интервала, согласно формуле (2.2), имеем:

$$p_i = \int_{x_i}^{x_i + \Delta x} f(x)dx,$$

а в случае нормального распределения $p_i$ равно [по формуле (2.4)]:

$$p_i = P\{x_i < x < x_i + \Delta x\} = \Phi\left(\frac{x_i + \Delta x - \bar{x}}{S}\right) - \Phi\left(\frac{x_i - \bar{x}}{S}\right).$$

На каждом интервале величина, характеризующая относительное отклонение выборочного распределения от гипотетического отклонения, определяется величиной

$$\frac{(n_i - np_i)^2}{np_i}.$$

Сумма этих отклонений по всем $k$ интервалам является случайной величиной, имеющей распределение $\chi^2$:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

с числом степеней свободы $f$:

$$f = k - c - 1,$$

где $c$ – число параметров предполагаемого генерального гипотетического распределения $f(x)$, точечные оценки которых вычисляются по выборке.

Так, в случае проверки гипотезы о нормальном распределении экспериментальных данных $c = 2$.

Гипотеза о проверяемом законе распределения принимается на принятом уровне значимости, если выполняется следующее неравенство:

$$\chi^2 < \chi^2_{1-\alpha,f},$$

где $\chi^2_{1-\alpha,f}$ – квантиль распределения $\chi^2$.

Корректное использование критерия Пирсона допустимо для выборок достаточно большого объема ($n \geq 50 \ldots 150$). Кроме того, количество результатов в каждом из интервалов должно быть не менее пяти. Если $n_i < 5$, следует объединить соседние интервалы.

### *3.7.3. Критерий Колмогорова*

Критерий Колмогорова используется для проверки гипотезы $H_0$ о том, что распределение генеральной совокупности описывается функцией распределения $F(x)$. Критерием проверки выдвинутой гипотезы служит статистика $D_n$:

$$D_n = \max_{-\infty \le x \le \infty} |F_n(x) - F(x)|, \qquad (3.20)$$

где $F_n(x)$ – эмпирическая функция распределения, построенная по данным ранжированной по возрастанию выборки, т. е. по выборке

$$x_{\min} = x_1 \le x_2 \le \cdots \le x_n = x_{\max}. \qquad (3.21)$$

Функция распределения $F_n(x)$ равна:

$$F_n(x) = \begin{cases} 0 \text{ при } x < x_{\min}, \\ i/n \text{ при } x_i \le x \le x_{i+1} \ (i = 1, \dots, n-1), \\ 1 \text{ при } x \ge x_{\max}. \end{cases}$$

График эмпирической функции $F(x)$, приведенный на рис. 3.2., в промежутках между соседними членами вариационного ряда (3.21) сохраняет постоянное значение, а в точках $x = x_i$ претерпевает разрыв, скачком возрастая на величину $1 / n$ (в случае совпадения $m$ наблюдений – на величину $m / n$).

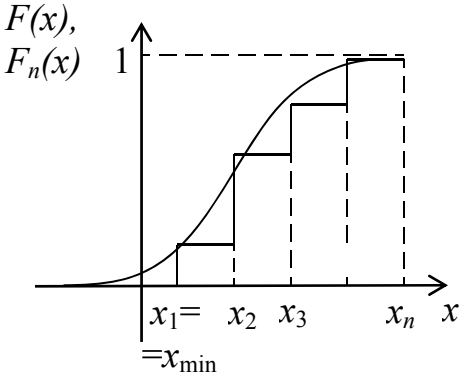


Рис. 3.2. График эмпирической функции $F(x)$
к проверке гипотезы согласия по критерию Колмогорова

Значение статистики, вычисленное по данным выборки по формуле (3.20), сравнивается с критическим значением критерия Колмогорова $K_{n,\alpha}$ на уровне значимости α [2; 6]. Выдвинутая гипотеза принимается, если выполняется неравенство $D_n < K_{n,\alpha}$ .

## 4. ОСНОВЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА

### 4.1. Задача корреляционного анализа. Генеральный коэффициент парной корреляции

Основная задача корреляционного анализа связана с выявлением зависимости между случайными переменами ХТО. Зависимость между двумя случайными величинами называется корреляционной, если с изме-

нением одной случайной величины происходит изменение некоторой числовой характеристики другой случайной величины [2; 5].

Пусть $x$ и $y$ – две случайные переменные, имеющие совместное нормальное распределение. Дисперсия суммы этих переменных $D(x+y)$ равна:

$$D(x+y) = M\{[(x+y) - M(x+y)]^2\} = M\left\{\left[(x-m_x) + \left(y-m_y\right)\right]^2\right\} =$$

$$= M\{(x-m_x)^2\} + M\left\{\left(y-m_y\right)^2\right\} + 2M\left\{(x-m_x)\left(y-m_y\right)\right\} =$$

$$= D_x + D_y + 2K_{xy},$$

где $K_{xy}$ – второй смешанный центральный момент (корреляционный момент) или ***ковариация*** случайных переменных $x$ и $y$.

Ковариация случайных переменных $K_{xy}$ равна:

$$K_{xy} = M\left\{(x-m_x)(y-m_y)\right\} = Cov(x,y).$$

Если случайные переменные $x$ и $y$ независимы, то $D(x+y) = D(x) + D(y)$, следовательно, $\mathrm{Cov}(x,y) = 0$ и величина ковариации может служить мерой тесноты связи между переменными $x$ и $y$. Но поскольку значение ковариации зависит еще и от вариации переменных $x$ и $y$ относительно их математических ожиданий, то для характеристики статистической зависимости между ними используется величина ρ:

$$\rho = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y} = M\left\{\frac{x-m_x}{\sigma_x} \cdot \frac{y-m_y}{\sigma_y}\right\},$$

где ρ – ***генеральный коэффициент парной корреляции***.

Для независимых случайных величин $\rho = 0$.

Если условие нормального распределения пары $x$ и $y$ не выполняется, то на основании равенства $\rho = 0$ можно делать заключение только о некоррелированности переменных, т. к. между ними может существовать нелинейная зависимость, т. е. коэффициент корреляции характеризует только тесноту линейной зависимости между $x$ и $y$.

Наглядное отображение зависимости двух случайных величин можно получить по виду ***поля корреляции*** (рис. 4.1.). Для его построения на координатной плоскости $x - y$ необходимо нанести $n$ точек с координатами $x_i$ и $y_i$, соответствующими результатам -го опыта в выборке объемом $n$.

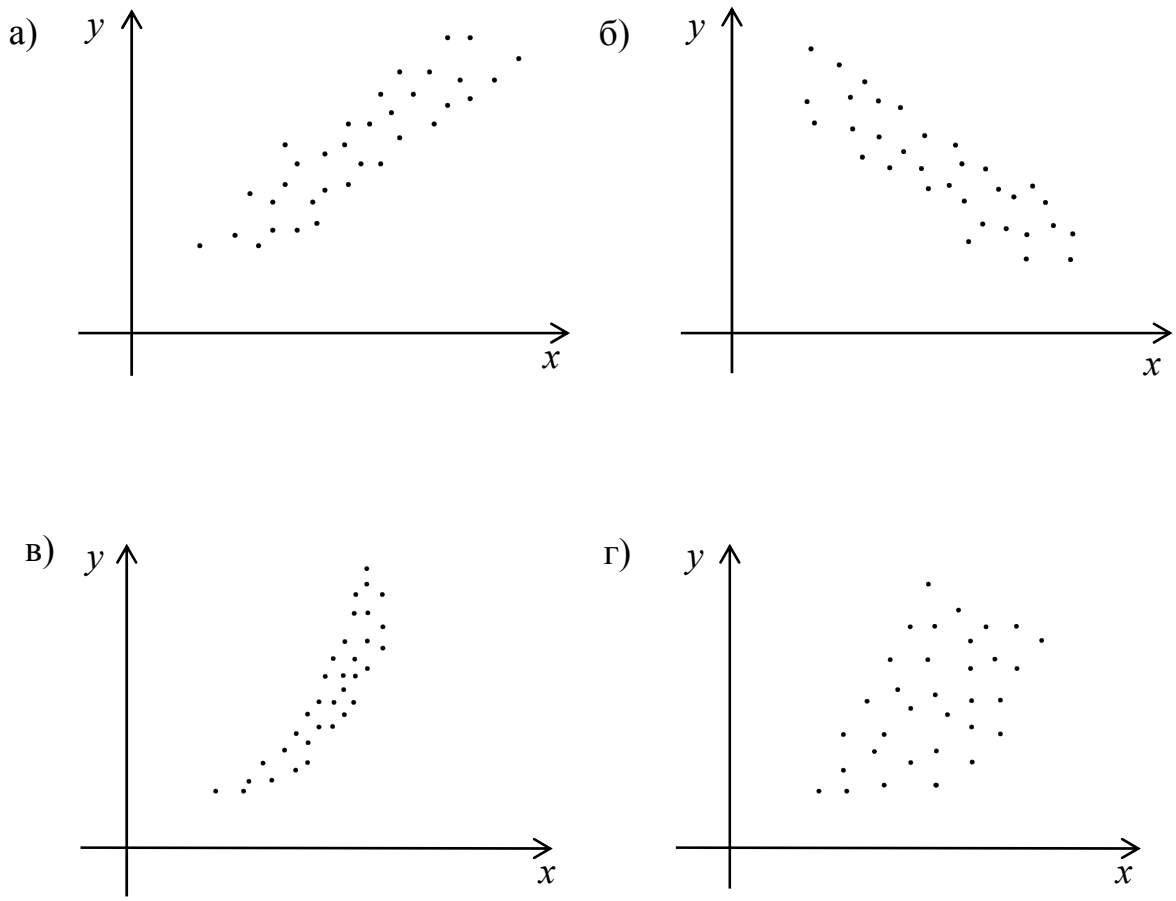


Рис. 4.1. Вид поля корреляции переменных $x$ и $y$:
а – наличие сильной положительной корреляции;
б – наличие сильной отрицательной корреляции;
в – нелинейная стохастическая зависимость;
г – отсутствие корреляции

## 4.2. Выборочный парный коэффициент корреляции

Оценкой генерального коэффициента корреляции ρ является ***выборочный коэффициент парной корреляции*** $r_{xy}$ (коэффициент корреляции Пирсона), величина которого вычисляется следующим образом:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}. \qquad (4.1)$$

Выражение (4.1) может быть преобразовано к более удобному для вычисления виду:

$$r_{xy} = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}. \quad (4.2)$$

Выборочный коэффициент парной корреляции $r_{xy}$, как и генеральный коэффициент ρ, обладает следующими свойствами:

1) Коэффициент $r_{xy}$ может принимать значения на отрезке $[-1; 1]$, т. е. $|r_{xy}| \leq 1$. Значения $r_{xy} = \pm 1$ соответствуют функциональной связи между переменными.

2) Коэффициент $r_{xy}$ не изменяется от прибавления к $x$ или $y$ любых неслучайных величин или от умножения $x$ или $y$ на положительное число. Указанные преобразования являются линейными и связаны лишь с изменением масштаба или начала отсчета случайных величин $x$ и $y$. Поэтому для нормированных переменных $x_0$ и $y_0$, равных:

$$x_0 = \frac{x - \bar{x}}{s_x},$$

$$y_0 = \frac{y - \bar{y}}{s_y}$$

справедливо равенство

$$r_{x_0 y_0} = r_{xy}.$$

3) $r_{xy} = r_{yx}$ [свойство следует из выражения (4.1)];

4) $r_{xx} = 1$ [свойство следует из выражения (4.1)].

Отличие выборочного коэффициента корреляции $r_{xy}$ от нуля еще не означает, что случайные величины находятся в стохастической зависимости. Следует проверить значимость выборочного коэффициента парной корреляции, т. е. установить, достаточно ли величины $r_{xy}$ для обоснованного вывода о наличии линейной корреляционный связи. С этой целью проверяют нулевую гипотезу $H_0$: $\rho = 0$, для чего вычисляют статистику $t$:

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}}, \quad (4.3)$$

которая имеет распределение Стьюдента с числом степеней свободы $f = n - 2$.

Если абсолютная величина статистики, рассчитанная по формуле (4.3), превышает критическое значение гипотезы $t_{f,1-\frac{\alpha}{2}}$, соответствующее выбранному уровню значимости α, т. е. выполняется неравенство

$$|t| \geq t_{f,1-\frac{\alpha}{2}},$$

то нулевую гипотезу следует отвергнуть и признать, что переменные $x$ и $y$ статистически зависимы. При $|t| < t_{f,1-\frac{\alpha}{2}}$ нет основания отвергать нулевую гипотезу, а отличие выборочного коэффициента парной корреляции от нуля обусловлено только случайным характером данной выборки.

Если выборочный коэффициент корреляции значим, то можно построить доверительный интервал для генерального коэффициента ρ, для чего вводят вспомогательную переменную $z$, которая связана с коэффициентом $r_{xy}$ с помощью преобразования Фишера:

$$z = \frac{1}{2}\ln\frac{1+r_{xy}}{1-r_{xy}}. \quad (4.4)$$

Обратное преобразование:

$$r_{xy} = \text{th}z = \frac{e^{2z}-1}{e^{2z}+1}, \quad (4.5)$$

где $\text{th}z$ – гиперболический тангенс $z$.

Распределение $z$ близко к нормальному распределению с параметрами

$$m_z = \frac{1}{2}\ln\frac{1+\rho}{1-\rho},$$

$$\sigma_z = \frac{1}{\sqrt{n-3}}.$$

Задавшись доверительной вероятностью $p$ и вычислив согласно формуле (4.4) величину $z$, соответствующую выборочному значению $r_{xy}$, можно определить границы доверительного интервала $(z_1; z_2)$ для $m_z$:

$$z - \frac{U_p}{\sqrt{n-3}} < m_z < z + \frac{U_p}{\sqrt{n-3}},$$

где $U_p$ – квантиль нормального распределения.

По найденным значениям $z_1$ и $z_2$ в соответствии с формулой (4.5) определяются границы доверительного интервала для генерального коэффициента ρ.

## 4.3. Частный и множественный коэффициенты корреляции

Если выходная переменная ХТО $y$ зависит от нескольких факторов $x_i$ ($i$=1, …, $k$), то использование парного коэффициента корреляции $r_{yx_i}$ для оценки тесноты связи между переменной $y$ и выбранным фактором $x_i$ недопустимо, т. к. изменение величины функции отклика обусловлено вариацией всех факторов.

В этом случае необходимо ориентироваться на значения ***частных коэффициентов корреляции***. Последние позволяют оценить тесноту связи между переменной $y$ и выбранным фактором при условии, что влияние остальных факторов на выходную переменную исключено.

Для расчета частных коэффициентов корреляции в случае $k$ факторов необходимо воспользоваться корреляционной матрицей $[r_{ij}]$, составленной из парных коэффициентов корреляции:

$$[r_{ij}] = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1k} & r_{1y} \\ r_{21} & r_{22} & \dots & r_{2k} & r_{2y} \\ r_{31} & r_{32} & \dots & r_{3k} & r_{3y} \\ \dots & \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{yk} & r_{ky} \\ r_{y1} & r_{y2} & \dots & r_{yk} & r_{yy} \end{pmatrix},$$

где для упрощения записи принято, что $r_{ij} = r_{x_i x_j}$, $r_{iy} = r_{x_i y}$.

Коэффициент частной корреляции между $y$ и $x_i$ вычисляется по формуле

$$r_{\frac{x_i y}{x_1}, x_2,\dots,x_{i-1},x_{i+1},\dots,x_k} = \frac{-C_{iy}}{\sqrt{C_{ii}C_{yy}}}, \qquad (4.6)$$

где $C_{iy}$ – алгебраическое дополнение для элемента $r_{x_i y,}$ стоящего в $i$-й строке последнего столбца; $C_{ii}$ – алгебраическое дополнение диагонального элемента $r_{x_i x_i}$, стоящего в -й строке; $C_{yy}$ – алгебраическое дополнение элемента $r_{yy}$.

Индекс в записи $r_{x_i y/x_1,x_2,\dots,x_{i-1},x_{i+1},\dots,x_k}$ указывает на фактор, для которого рассчитывается частный коэффициент корреляции с функцией отклика, при этом влияние остальных факторов (перечисленных в индексе после черты) на выходную переменную ХТО исключено.

В случае двух факторов $x_1$ и $x_2$ корреляционная матрица имеет следующий вид:

$$[r_{ij}] = \begin{pmatrix} 1 & r_{12} & r_{1y} \\ r_{21} & 1 & r_{2y} \\ r_{y1} & r_{y2} & 1 \end{pmatrix},$$

а частные коэффициенты корреляции, согласно формуле (4.6), равны:

$$r_{\underline{yx_1}}_{x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}},$$

$$r_{\underline{yx_2}}_{x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{x_1x_2}^2\right)}}.$$

Коэффициент $r_{\underline{yx_1}}_{x_2}$ отражает уровень связи между $y$ и $x_1$, когда влияние фактора $x_2$ исключено; коэффициент $r_{\underline{yx_2}}_{x_1}$ характеризует зависимость между $y$ и $x_2$ при условии, что значение $x_1$ неизменно.

Мерой линейной зависимости выходной переменной ХТО от всех входных воздействий является ***множественный коэффициент детерминации*** $R^2_{y,x_1,x_2,\ldots,x_k}$, равный:

$$R^2_{y,x_1,x_2,\ldots,x_k} = 1 - \frac{|r_{xy}|}{C_{yy}},$$

где $|r_{xy}|$ – определитель корреляционной матрицы.

Квадратный корень из $R^2_{y,x_1,x_{2,}\ldots,x_k}$ называется ***коэффициентом множественной корреляции***.

# 5. ОСНОВЫ РЕГРЕССИОННОГО АНАЛИЗА

## 5.1. Построение статистических моделей химико-технологических объектов на основе результатов пассивного эксперимента

Основная задача регрессионного анализа – построение стохастической регрессионной модели ХТО, адекватно описывающей зависимость между его выходными и входными переменными. В зависимости от числа факторов различают парный (одномерный) регрессионный анализ $y = f(x)$ и множественный регрессионный анализ $y = f(x_1, x_2, \ldots, x_k)$.

При осуществлении регрессионного анализа предполагается выполнение следующих условий:

1) В ходе проведения опыта значение каждого из факторов фиксируется на определенном неизменном уровне, а если вариация фактора и имеет место, то ее влияние на величину функции отклика незначительно.

2) В случае парного регрессионного анализа фактор $x$ может принимать $n$ различных значений $x_i$ ($i$=1, …, $n$), а в случае построения множественной регрессионной модели возможно $n$ различных сочетаний значений $k$ факторов.

3) Для каждого значения фактора (сочетания факторов) опыт повторяется $m_i$ раз (выполняется серия из $m_i$ параллельных опытов). Если дублирование опытов не предусмотрено, полагаем $m_i$=1. В дальнейшем под $y_{ij}$ подразумевается значение функции отклика в $j$-м параллельном опыте из серии, соответствующей значению фактора, зафиксированного на уровне $x_i$ ($j$=1, …, $m_i$).

4) Результаты каждой серии параллельных опытов представляют собой реализации независимых случайных величин, имеющих нормальное распределение. Для каждой серии вычисляются точечные оценки математического ожидания и дисперсии:

$$\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij},$$

$$s_{yi}^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2,$$

где $i$=1, …, $n$.

5) Выборочные дисперсии $s_{yi}^2$, вычисленные по указанным сериям, должны быть однородными. После принятия гипотезы об однородности выборочных дисперсий $s_{yi}^2$ вычисляется дисперсия воспроизводимости эксперимента $s_{\text{воспр}}^2$. Если дублирование экспериментов при фиксированных значениях $x_i$ не производилось, то для определения дисперсии $s_{\text{воспр}}^2$ проводят специальную серию опытов.

Далее рассмотрим методику построения регрессионной модели, включающую в себя следующие основные разделы:

1) выбор структуры модели;

2) определение коэффициентов модели и проверку их значимости;

3) оценку адекватности полученной модели.

### *5.1.1. Выбор структуры модели*

Среди перечисленных выше разделов данный раздел наименее формализован. Неверный выбор структуры – наиболее вероятная причина неадекватности модели. При построении одномерной модели некоторое

представление о структуре регрессионной зависимости может дать вид поля корреляции (см. рис. 4.1.). Обычно, если это не противоречит теоретическим сведениям о функционировании ХТО и подтверждается практическим опытом, накопленным в процессе эксплуатации объекта, первоначально рассматриваются наиболее простые структуры. К их числу относится парная линейная модель ХТО, которая имеет следующий вид:

$$\hat{y} = \beta_0 + \beta_1 x, \qquad (5.1)$$

где $\beta_0$, $\beta_1$ – генеральные коэффициенты модели.

Если регрессия (5.1) окажется неадекватной, то следует перейти к более сложным одномерным моделям, например:

1) к полиномиальной регрессии 2-го порядка $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$;
2) к гиперболической регрессионной модели $\hat{y} = \beta_0 + \frac{\beta_1}{x}$;
3) к показательной регрессионной зависимости $\hat{y} = \beta_0 + \beta_1 \beta_2^x$.

Но, что более вероятно, причина неадекватности парной регрессионной зависимости обусловлена тем, что процессы, протекающие в промышленных ХТО, определяются влиянием достаточно большого числа факторов, и для построения адекватных моделей таких объектов следует перейти от одномерных регрессионных зависимостей к множественным.

### *5.1.2. Определение коэффициентов модели и проверка их значимости*

При построении как парной, так и множественной регрессионных зависимостей исходной информацией для расчета являются результаты эксперимента, реализованного на ХТО. В случае парной регрессии эта информация представлена в виде $n$ пар значений $x_i$ и $y_i$, а при выполнении параллельных опытов – в виде $x_i$ и $\bar{y}_i$.

Поскольку экспериментальные данные представляют собой случайные выборки из генеральной совокупности, то полученные в результате расчета коэффициенты представляют собой выборочные оценки генеральных коэффициентов $\beta_i$, ($i$=0, …, $l$), где $l$ – число коэффициентов модели. Выборочные коэффициенты модели, в отличие от генеральных коэффициентов, обозначаются как $b_i$ ($i$=0, …, $l$).

Наиболее распространенным методом определения коэффициентов $b_i$ является ***метод наименьших квадратов*** (МНК). Согласно МНК, наилучшими считаются такие коэффициенты, которым соответствует минимальная сумма квадратов отклонений результатов экспериментов $y_i$ от значений функции отклика $\hat{y}_i$, рассчитанных по модели, т. е.

$$U(b_0, b_1, b_2, .. b_l) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min. \qquad (5.2)$$

В случае построении парной линейной модели (5.1) критерию (5.2) будет соответствовать условие

$$U(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \to \min. \qquad (5.3)$$

Для нахождения экстремума (минимума) выражения $U(b_0, b_1)$, являющегося функцией двух переменных $b_0$ и $b_1$, необходимо решить следующую систему уравнений:

$$\left.\begin{aligned} \frac{\partial U(b_0, b_1)}{\partial b_0} = 0, \\ \frac{\partial U(b_0, b_1)}{\partial b_1} = 0, \end{aligned}\right\}$$

которая после вычисления частных производных от выражения (5.3) и преобразования принимает следующий вид:

$$\left.\begin{aligned} nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{aligned}\right\} \qquad (5.4)$$

Выражения для коэффициентов $b_0$ и $b_1$ получают, решая систему уравнений (5.4), которая в регрессионном анализе называется ***системой нормальных уравнений***:

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

С учетом выражения (4.2) для выборочного коэффициента корреляции $r_{xy}$ коэффициент $b_1$ может быть рассчитан по формуле

$$b_1 = r_{xy}\frac{s_y}{s_x}.$$

При этом уравнение парной линейной модели ХТО может быть записано в одной из следующих форм:

$$\hat{y} = b_0 + b_1 x,$$

$$\hat{y} = \bar{y} + b_1(x - \bar{x}),$$

$$\hat{y} = \bar{y} + r_{xy}\frac{s_y}{s_x}(x - \bar{x}).$$

После того как коэффициенты уравнения регрессии определены, необходимо оценить их значимость. Проверка значимости коэффициентов производится на основе проверки статистической гипотезы $H_0: \beta_j = 0$. Проверка осуществляется с помощью критерия Стьюдента:

$$t_j = \frac{|b_j|}{s_{b_j}}, \tag{5.5}$$

где $s_{b_j}$ – среднеквадратичное отклонение выборочного коэффициента $b_j$.

Для парной линейной модели ХТО [см. формулу (5.1)] расчет значений среднеквадратических отклонений $s_{b_0}$ и $s_{b_1}$ производится по следующим формулам:

$$S_{b_0} = \sqrt{\frac{S^2_{\text{воспр}} \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}},$$

$$S_{b_1} = \sqrt{\frac{S^2_{\text{воспр}} n}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}.$$

Критическое значение критерия $t_{f,1-\alpha}$ принимается на уровне значимости α для числа степеней свободы $f = f_{\text{воспр}} = \sum_{i=1}^{n} m_i - n$.

Если $t_j < t_{f,1-\alpha}$, то коэффициент $b_j$ признается незначимым и исключается из уравнения регрессии. Оставшиеся коэффициенты модели необходимо пересчитать, поскольку они взаимосвязаны (закоррелированы) друг с другом.

### *5.1.3. Оценка адекватности полученной модели*

На этом этапе необходимо установить, соответствует ли полученная регрессионная зависимость тем результатам эксперимента, по которым она была построена.

Расхождение между результатами экспериментов $y_i$ и значениями функции отклика, рассчитанными по модели $\hat{y}_i$, обусловлено двумя основными причинами. Это, во-первых, неадекватность полученной модели, а во-вторых, искажение результатов эксперимента, вызванное влиянием неучтенных, случайным образом изменяющихся переменных. Суммарный эффект от действий обеих причин можно оценить по величине остаточной дисперсии:

$$s^2_{\text{ост}} = \frac{ss_{\text{ост}}}{f_{\text{ост}}}, \tag{5.6}$$

где

$$ss_{\text{ост}} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - \hat{y}_i)^2,$$

$$f_{\text{ост}} = \sum_{i=1}^{n} m_i - l,$$

где $m_i$ – объем $i$-й серии параллельных опытов; $l$ – число коэффициентов модели.

Для проверки адекватности модели сумму квадратов остаточной дисперсии $ss_{\text{ост}}$ необходимо представить в виде двух составляющих, первая из которых $(ss_{\text{ад}})$ определяется неточностью модели, а вторая $(ss_{\text{воспр}})$ – влиянием на ХТО неучтенных переменных, т. е. $ss_{\text{ост}}$ необходимо представить в следующем виде:

$$ss_{\text{ост}} = ss_{\text{ад}} + ss_{\text{воспр}}. \tag{5.7}$$

Сумма квадратов $ss_{\text{воспр}}$, характеризующая степень воспроизводимости результатов эксперимента, определяется разбросом случайных значений выходной переменной $y_{ij}$ относительно выборочных средних $\bar{y}_i$:

$$ss_{\text{воспр}} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2. \tag{5.8}$$

С суммой квадратов $ss_{\text{воспр}}$ связана величина дисперсии воспроизводимости $s^2_{\text{воспр}}$:

$$s^2_{\text{воспр}} = \frac{ss_{\text{воспр}}}{f_{\text{воспр}}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^{n} m_i - n}, \tag{5.9}$$

где число степеней свободы равно: $f_{\text{воспр}} = \sum_{i=1}^{n} m_i - n$.

Согласно выражениям (5.6) – (5.9), сумма квадратов $ss_{\text{ад}}$ равна:

$$ss_{\text{ад}} = ss_{\text{ост}} - ss_{\text{воспр}} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - \hat{y}_i)^2 - \sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2.$$

Зная величину $ss_{\text{ад}}$, можно определить ***дисперсию адекватности***, характеризующую уровень адекватность полученной модели:

$$s_{\text{ад}}^2 = \frac{ss_{\text{ад}}}{f_{\text{ад}}}. \qquad (5.10)$$

Число степеней свободы дисперсии адекватности равно:

$$f_{\text{ад}} = f_{\text{ост}} - f_{\text{воспр}} = \left(\sum_{i=1}^{n} m_i - l\right) - \left(\sum_{i=1}^{n} m_i - n\right) = n - l.$$

В случае одномерной линейной модели $l = 2$.

Для принятия гипотезы о том, что полученная модель является адекватной, необходимо проверить значимость различия между дисперсией адекватности $s_{\text{ад}}^2$ и дисперсией воспроизводимости $s_{\text{воспр}}^2$.

Если серии параллельных опытов имеют одинаковый объем $m$, то расчетные формулы (5.9) и (5.10) принимают следующий вид:

$$s_{\text{воспр}}^2 = \frac{\sum_{i=1}^{n}\sum_{i=1}^{m}(y_{ij} - \bar{y}_i)^2}{nm - n},$$

$$s_{\text{ад}}^2 = \frac{m\sum_{i=1}^{n}(\bar{y}_i - \hat{y}_i)^2}{n - l}.$$

В случае, когда дублирование экспериментов не производится, а для определенности дисперсии воспроизводимости ставится специальная серия опытов объемом $m_0$, расчет дисперсий $s_{\text{воспр}}^2$ и $s_{\text{ад}}^2$ по результатам этой серии $y_i^0$ ($i$=0, …, $m_0$) производится следующим образом:

$$s_{\text{воспр}}^2 = \frac{\sum_{i=1}^{m_0}(y_i^0 - \overline{y^0})^2}{m_0 - 1},$$

$$s_{\text{ост}}^2 = s_{\text{ад}}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - l}.$$

Проверка значимости различия между дисперсией адекватности и дисперсией воспроизводимости осуществляется по критерию Фишера, согласно которому стохастическая модель считается адекватной, если на принятом уровне значимости α выполняется неравенство

$$F = \frac{s^2_{\text{ад}}}{s^2_{\text{воспр}}} < F_{1-\alpha, f_{\text{ад}}, f_{\text{воспр}}}.$$

Если парная линейная модель ХТО [см. формулу (5.1)] оказалась неадекватной, то можно перейти к рассмотрению нелинейной зависимости функции отклика от фактора $x$: например, ввести в модель член $x^2$, т. е. увеличить порядок полиномиальной модели. В качестве нелинейных парных моделей могут быть рассмотрены гиперболические, показательные или иные модели, которые, в отличие от полиномиальных, относятся к классу моделей ***нелинейных по параметрам*** [1; 6]. Для расчета коэффициентов таких моделей метод наименьших квадратов может быть применен только после необходимых преобразований переменных, позволяющих линеаризовать исходную модель.

Принципиально другой подход к построению адекватной модели ХТО связан с переходом от одномерный регрессии $\hat{y} = f(x)$ к множественной регрессии $\hat{y} = f(x_1, x_2, \dots, x_k)$, согласно которому в модель вводятся нескольких факторов, значимо влияющих на величину функции отклика $y$.

## 5.2. Нелинейная парная регрессия

Повышая степень полинома в регрессивной модели, в некоторых случаях можно добиться ее адекватности. Переходя от полинома степени $m$ к полиному степени $(m + 1)$, необходимо пересчитывать коэффициенты уравнения регрессии.

Для новой модели необходимо рассчитать величину остаточной дисперсии:

$$s^2_{\text{ост}(m+1)} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - l},$$

где $l$ – число значимых коэффициентов модели.

Величину $s^2_{\text{ост}(m+1)}$ необходимо сравнить с остаточной дисперсией модели степени $m$. Если различие между ними незначимо, т. е.

$$F = \frac{s^2_{\text{ост}(m)}}{s^2_{\text{ост}(m+1)}} < F_{1-\alpha, f_{(m)}, f_{(m+1)}},$$

то дальнейшее увеличение порядка полинома $y = f(x)$ не приведет к увеличению степени адекватности модели.

Для случая параболической парной регрессии $\hat{y} = b_0 + b_1 x + b_2 x^2$ система нормальных уравнений для расчета коэффициентов модели согласно МНК имеет следующий вид:

$$\left.\begin{aligned} b_0 n + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i\,, \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} x_i y_i\,, \\ b_0 \sum_{i=1}^{n} x_i^2 + b_1 \sum_{i=1}^{n} x_i^3 + b_2 \sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} x_i^2 y_i. \end{aligned}\right\} \qquad (5.11)$$

Структура системы нормальных уравнений сохранится при увеличении порядка регрессионной модели.

При описании экспериментальных данных могут быть использованы нелинейные зависимости, которые с помощью преобразования переменных сводятся к линейным. Коэффициенты этих линеаризованных моделей определяются с использованием МНК, после чего с помощью обратных преобразований рассчитываются коэффициенты нелинейных моделей.

Например, в случае показательной модели

$$\hat{y} = b_0 {b_1}^x \qquad (5.12)$$

линеаризация осуществляется с помощью преобразования выходной переменной $\hat{y}$ к новой переменной $\hat{z} = \ln \hat{y}$. Получаем линеаризованную модель:

$$\hat{z} = a_0 + a_1 x,$$

где $a_0 = \ln b_0$ ; $a_1 = \ln b_1$.

По коэффициентам $a_0$ и $a_1$, рассчитанным с использованием МНК, определяются коэффициенты исходной нелинейной модели [см. формулу (5.12)]:

$$b_0 = e^{a_0},$$

$$b_1 = e^{a_1}.$$

Для степенной модели $\hat{y} = b_0 x^{b_1}$ переход к линеаризованной зависимости $\hat{z} = a_0 + b_1 t$ осуществляется преобразованием экспериментальных данных по формулам

$$\hat{z} = \ln \hat{y},$$

$$t = \ln x.$$

Обратное преобразование коэффициентов: $b_0 = e^{a_0}$.

Для линеаризации гиперболической модели

$$\hat{y} = b_0 + \frac{b_1}{x}$$

с использованием преобразования входных переменных $t = 1/x$ осуществляется переход к линеаризованной модели $y = b_0 + b_1 t$.

### 5.3. Множественный линейный регрессионный анализ

При исследовании ХТО парные регрессионные зависимости $y = f(x)$ используются достаточно редко. Более распространенной является ситуация, в которой выходная переменная объекта $y$ зависит от $k$ факторов, где $k \geq 2$.

На первом этапе анализа целесообразно ограничится использованием линейной модели, пренебрегая эффектами взаимодействия факторов, т. е. членами $b_{ij}x_i x_j$ и квадратичными членами $b_{ii}x_i^2$. При этом задача регрессионного анализа сводится к построению модели

$$\hat{y} = b_0 x_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k, \qquad (5.13)$$

где $x_0$ – фиктивная переменная, равная единице.

Решение этой задачи с помощью МНК удобно проводить в матричной форме. С этой целью весь набор значений факторов, зафиксированных в $n$ опытах, оформляется в виде ***матрицы наблюдений*** (матрица независимых или входных переменных):

$$\boldsymbol{X} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & x_{13} & \dots\dots & x_{1k} \\ x_{20} & x_{21} & x_{22} & x_{23} & \dots\dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i0} & x_{i1} & x_{i2} & x_{i3} & \dots x_{ij} \dots & \dots \\ x_{n0} & x_{n1} & x_{n2} & x_{n3} & \dots\dots & x_{nk} \end{pmatrix},$$

где $x_{ij}$ – значение $j$-го фактора в -м опыте ($i = 1, \dots, n$; $j = 0, \dots, k$), т. е. размерность матрицы наблюдений равна $n(k+1)$ (ранее отмечено, что $x_{i0} = 1$).

Результаты эксперимента представляется в виде матрицы-столбца $\boldsymbol{Y}$ ($n$-мерного вектора наблюдений):

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_n \end{pmatrix},$$

где $y_i$ – значение выходной величины ХТО в -м опыте в случае, когда проведение параллельных опытов со значениями факторов, соответствующих $i$-ой сроке матрицы наблюдений, не предусмотрено, или среднее значение

выходных переменных в $i$-й серии параллельных опытов, если планом эксперимента предусмотрено дублирование каждой строки матрицы наблюдений.

Подлежащие определению значения коэффициентов модели образуют матрицу-столбец $\boldsymbol{B}$:

$$\boldsymbol{B} = \begin{pmatrix} b_0 \\ b_1 \\ b_j \\ b_k \end{pmatrix}.$$

Значения функции отклика $\hat{y}_i$ рассчитываются по модели (5.13):

$$\hat{\boldsymbol{Y}} = \boldsymbol{XB}$$

и представляются в виде матрицы-столбца:

$$\hat{\boldsymbol{Y}} = \begin{pmatrix} \hat{y}_1 \\ \dots \\ \hat{y}_i \\ \dots \\ \hat{y}_n \end{pmatrix}.$$

Критерий оптимальности выбора коэффициентов модели согласно МНК [см. формулу (5.2)] записывается в матричной форме следующим образом:

$$U(b_0, b_1, .. b_k) = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \left(\boldsymbol{Y}_i - \boldsymbol{XB}\right)^{\mathrm{T}} \left(\boldsymbol{Y}_i - \boldsymbol{XB}\right) \rightarrow \min.$$

Структура системы нормальных уравнений в случае множественной регрессии совпадает со структурой систем нормальных уравнений, используемых для расчета коэффициентов линейной и параболической парных регрессий, определяемых выражениями (5.4) и (5.11) соответственно. В матричной форме система нормальных уравнений записывается как

$$\boldsymbol{X}^T \boldsymbol{X}\, \boldsymbol{B} = \boldsymbol{X}^T \boldsymbol{Y},$$

а ее решение имеет следующий вид:

$$\boldsymbol{B} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}. \qquad (5.14)$$

Матрица $\boldsymbol{X}^T \boldsymbol{X}$ называется ***матрицей моментов***. Это квадратная симметричная матрица размерностью $(k + 1)$.

Матрица $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$, обратная матрице моментов, называется ***матрицей ошибок*** или ***ковариационной матрицей***. Для того чтобы матрица ошибок существовала, матрица моментов должна быть невырожденной, т. е.

$$det(\boldsymbol{X}^T \boldsymbol{X}) \neq 0.$$

Величина определителя матрицы $\boldsymbol{X}^T\boldsymbol{X}$ связана со ***степенью обусловленности*** системы нормальных уравнений. Если величина определителя $det(\boldsymbol{X}^T\boldsymbol{X})$ мала, то система оказывается плохо обусловленной. При этом небольшим изменениям матрицы наблюдений $\boldsymbol{X}$ соответствуют значительные изменения коэффициентов модели.

Поскольку коэффициенты регрессии взаимосвязаны, раздельно оценить значимость каждого из них невозможно. Поэтому после определения коэффициентов $b_j$ $(j = 0, \dots, k)$ вычисляются соответствующие им статистики $t_j$:

$$t_j = \frac{|b_j|}{s_{\text{воспр}}\sqrt{c_{jj}}},$$

где $c_{jj}$ –диагональный коэффициент ковариационной матрицы.

Фактор, для которого значение статистики $t_j$ оказалось наименьшим, исключается из уравнения регрессии. После этого расчет коэффициентов необходимо провести еще раз. Эта процедура повторяется до тех пор, пока будет значимо уменьшаться остаточная дисперсия $s^2_{\text{ост}}$ :

$$s^2_{\text{ост}} = \frac{ss_{\text{ост}}}{f_{\text{ост}}} = \frac{\sum_{i=1}^{n}(\bar{y}_i - \hat{y}_i)^2 = \left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right)^T\left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right)}{f_{\text{ост}}},$$

где вектор-столбец $\widehat{\boldsymbol{Y}}$ рассчитанных по модели значений выходной переменной вычисляется по формуле

$$\widehat{\boldsymbol{Y}} = \boldsymbol{XB}.$$

Число степеней свободы остаточной дисперсии $f_{\text{ост}}$ равно:

$$f_{\text{ост}} = n - l = n - k - 1,$$

где $l -$ число коэффициентов модели.

## 5.4. Построение статистических моделей химико-технологических объектов на основе активного эксперимента

### *5.4.1. Полный факторный эксперимент*

Активный эксперимент предполагает целенаправленный выбор значений факторов. По сравнению с пассивным экспериментом такой эксперимент оказывается более эффективным и экономичным, поскольку значительно сокращается объем экспериментальной работы. Недостаток активного эксперимента, предполагающего нарушение нормального хода техно-

логического процесса, связан с возможностью получения бракованного продукта и повреждением ХТО.

При планировании эксперимента каждый из факторов может принимать определенное число значений или уровней. Если число уровней равно $n$, а число факторов равно $k$, то ***полный факторный эксперимент*** (ПФЭ) предполагает выполнение $N$ опытов, где $N = n^k$. План такого эксперимента (матрица планирования) содержит совокупность всех $N$ возможных комбинаций факторов $k$, варьируемых на $n$ уровнях.

Наиболее распространен двухуровневый эксперимент, когда $n = 2$. При этом фактор $z_j$ принимает значения $z_j^{\max}$ и $z_j^{\min}$, соответствующие верхнему и нижнему уровням.

Величина $\Delta z_j = \frac{z_j^{\max} - z_j^{\min}}{2}$ представляет собой ***интервал варьирования*** для *j*-го фактора относительно ***основного*** или ***нулевого*** уровня $z_j^0$, равного:

$$z_j^0 = \frac{z_j^{\max} + z_j^{\min}}{2}.$$

Очевидно, что $z_j^{\min} = z_j^0 - z_j$ и $z_j^{\max} = z_j^0 + \Delta z_j$. Выбранные значения $z_j^0$ и $\Delta z_j$ определяют исследуемую область факторного пространства, в которой осуществляется постановка эксперимента.

Величина интервала $\Delta z_j$ должна быть достаточно большой, чтобы эффект от варьирования фактора не терялся на фоне случайных шумов ХТО. С другой стороны, завышение величины интервала вырывания затрудняет возможность адекватного описания объекта с помощью регрессионной модели.

В *k*-мерном факторном пространстве точка с координатами $(z_1^0, z_2^0, \dots, z_k^0)$ называется центром плана. Обработка результатов ПФЭ значительно упрощается, если от факторов $z_j$, записанных в натуральном масштабе, перейти к безразмерным переменным $x_j$ по формуле

$$x_j = \frac{z_j - z_j^0}{\Delta z_j}.$$

Тогда имеем:

$$x_j^{\max} = \frac{z_j^0 + \Delta z_j - z_j^0}{\Delta z_j} = 1,$$

$$x_j^{\max} = \frac{z_j^{\min} - z_j^0}{\Delta z_j} = \frac{-\Delta z_j}{\Delta z} = -1,$$

$$x_j^0 = \frac{z_j^0 - z_j^0}{\Delta z} = 0.$$

При этом матрица планирования принимает стандартный вид, при котором всем переменным на верхнем уровне соответствует значение +1, а на нижнем – значение -1. Иногда при заполнении матрицы планирования указывают только знаки уровней: плюс или минус.

***Пример 6:*** Рассмотрим пример заполнения матрицы планирования эксперимента, по результатам которого строится регрессионная зависимость степени разложения полигалита азотной кислоты [1].

Полигалит – минерал класса сульфатов ($K_2Ca_2Mg(SO_4)_4*2H_2O$) – сырье для производства минеральных удобрений.

В качестве факторов выбраны следующие параметры: $z_1$ –температура процесса, $^{\circ}$C; $z_2$ – продолжительность взаимодействия реагентов, мин; $z_3$ – концентрация азотной кислоты, масс. %.

В качестве выходных переменных рассматриваются следующие параметры: $y_1$ – степень извлечения $K_2O$, масс. %; $y_2$ – степень извлечения $MgO$, масс. %.

Нулевой уровень плана (центр плана) и интервалы варьирования:

$z_1^0 = 30\ ^{\circ}$C; $\Delta z_1 = 6\ ^{\circ}$C;
$z_2^0 = 14$ мин; $\Delta z_2 = 3$ мин;
$z_3^0 = 12{,}5$ %; $\Delta z_3 = 5$ %.

Число опытов (число возможных комбинаций уровней факторов) $N$ для ПФЭ равно $2^k$ , что в случае трех факторов равняется: $N = 2^3 = 8$. План эксперимента при записи факторов в натуральном масштабе приведен в табл. 5.1.

Таблица 5.1

План эксперимента при записи факторов в натуральном масштабе

| Номер опыта | $z_1$ | $z_2$ | $z_3$ | Номер опыта | $z_1$ | $z_2$ | $z_3$ |
|---|---|---|---|---|---|---|---|
| 1 | 24 | 11 | 7,5 | 5 | 24 | 11 | 17,5 |
| 2 | 36 | 11 | 7,5 | 6 | 36 | 11 | 17,5 |
| 3 | 24 | 17 | 7,5 | 7 | 24 | 17 | 17,5 |
| 4 | 36 | 17 | 7,5 | 8 | 36 | 17 | 17,5 |

Для рассматриваемого ПФЭ матрица планирования, записанная в безразмерных переменных с дополнительным столбцом фиктивного фактора $x_0$, равного единице, имеет вид, представленный в табл. 5.2.

Таблица 5.2

Матрица планирования ПФЭ $2^3$

| Номер опыта | $x_0$ | $x_1$ | $x_2$ | $x_3$ | Номер опыта | $x_0$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | +1 | –1 | –1 | –1 | 5 | +1 | –1 | –1 | +1 |
| 2 | +1 | +1 | –1 | –1 | 6 | +1 | +1 | –1 | +1 |
| 3 | +1 | –1 | +1 | –1 | 7 | +1 | –1 | +1 | +1 |
| 4 | +1 | +1 | +1 | –1 | 8 | +1 | +1 | +1 | +1 |

При составлении матрицы планирования пользуются следующим правилом: каждый последующий фактор (столбец) меняет знак вдвое реже, чем предыдущий.

Важными понятиями теории планирования эксперимента является рандомизация и число степеней свободы.

Рандомизацией называется любая процедура, обеспечивающая случайный порядок проведения эксперимента. Она позволяет исключить систематические воздействия неконтролируемых факторов. Например, в случае значительного воздействия неучтенного фактора при выполнении второй половины плана влияние фактора $x_3$ в модели будет искажено. При случайном порядке проведения экспериментов этого не произойдет.

Понятие числа степеней свободы использовалось при расчете дисперсии по формуле (3.3) и проверке статистических гипотез. Применительно к планированию эксперимента числом степеней свободы называют разность между числом экспериментов $N$ и числом налагаемых связей $l$, т. е. числом коэффициентов модели, рассчитанных по результатам эксперимента. Для рассматриваемого ПФЭ $2^3$ при построении линейной модели

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

число степеней свободы $f$ равно:

$$f = N - l = 8 - 4 = 4 > 0.$$

При $f > 0$ план эксперимента называется ***ненасыщенным***. Используя результаты эксперимента, проведенного по такому плану, можно рассчитать все коэффициенты модели и провести проверку ее адекватности.

В случае построения модели вида

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3$$

число степеней свободы $f$ равно: $f = 8 - 8 = 0$.

При $f = 0$ план становится ***насыщенным***. В этом случае возможна оценка всех параметров модели, но проверку адекватности провести нельзя.

В случае построения модели вида

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 +$$

$$+b_{11}x_1^2 + b_{22}x_2^2 + +b_{33}x_3^2$$

число степеней свободы $f < 0$, что соответствует ***сверхнасыщенным*** планам, при которых возможна оценка только части коэффициентов модели.

Матрица планирования ПФЭ обладает следующими свойствами:

1) Для любого фактора, кроме нулевого, сумма значений его уровней во всех опытах равна нулю:

$$\sum_{i=1}^{n} x_{ij} = 0,$$

где $j = 1, \dots, k$.

2) Для любого фактора сумма значений квадратов его уровней во всех опытах равна числу опытов в матрице планирования:

$$\sum_{i=1}^{n} x_{ij}^2 = N,$$

где $j = 1, \dots, k$.

3) Для любых двух факторов $x_s$ и $x_h$ сумма произведений значений уровней во всех опытах равна нулю:

$$\sum_{i=1}^{n} x_{is}x_{ih} = 0.$$

Данное свойство можно также сформулировать по-другому: скалярное произведение матриц-столбцов $x_s$ и $x_h$ равно нулю. При выполнении этого условия матрица планирования является ***ортогональной***. Если матрица $\boldsymbol{X}$ ортогональна, то соответствующая ей матрица моментов $(\boldsymbol{X}^T\boldsymbol{X})$ становится диагональной, т. е. все ее недиагональные элементы равны нулю, а элементы главной диагонали равны $N$.

Матрица ошибок (ковариационная матрица) $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, соответствующая матрице моментов, также является диагональной и все ее диагональные элементы равны $1/N$. При выполнении ПФЭ, являющегося частным случаем множественной регрессии, матрица-столбец коэффициентов модели может быть вычислена в по формуле (5.14):

$$\boldsymbol{B} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\ .$$

С учетом специфики ковариационной матрицы, связанной с ортогональностью матрицы планирования, расчет коэффициентов модели существенно упрощается. Действительно, для линейной модели с тремя факторами ковариационная матрица равна:

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{8} & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 \\ 0 & 0 & 0 & \frac{1}{8} \end{pmatrix}.$$

Скалярное произведение $\boldsymbol{X}^T\boldsymbol{Y}$ равно:

$$\begin{pmatrix} x_{10} & x_{20} & \cdots & x_{N0} \\ x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{21} & \cdots & x_{N2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{Nk} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} =$$

$$= \begin{pmatrix} x_{10}y_1 + x_{20}y_2 + \cdots + x_{N0}y_N \\ x_{11}y_1 + x_{21}y_2 + \cdots + x_{N1}y_N \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{1k}y_1 + x_{2k}y_2 + \cdots + x_{Nk}y_N \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_{i0}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{pmatrix}.$$

Тогда, согласно формуле (5.14), получаем:

$$\boldsymbol{B} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{8} & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 \\ 0 & 0 & 0 & \frac{1}{8} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{8} x_{i0}y_i \\ \sum_{i=1}^{8} x_{i1}y_i \\ \sum_{i=1}^{8} x_{i2}y_i \\ \sum_{i=1}^{8} x_{i3}y_i \end{pmatrix} = \frac{1}{8} \begin{pmatrix} \sum_{i=1}^{8} x_{i0}y_i \\ \sum_{i=1}^{8} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{8} x_{i3}y_i \end{pmatrix}.$$

Следовательно, любой коэффициент модели $b_j$ равен скалярному произведению столбца $x_j$ на матрицу-столбец наблюдений $y_j$,делимому на число опытов *N*:

$$b_i = \frac{1}{N}\sum_{i=1}^{n} x_{ij} y_i.$$

Если для рассматриваемого примера по результатам эксперимента получены следующие значения степени извлечения в жидкую фазу $K_2O_2$, масс. %:

$$y = \begin{pmatrix} 93 \\ 65 \\ 78 \\ 94 \\ 80 \\ 63 \\ 70 \\ 81 \end{pmatrix},$$

то значение коэффициента $b_0$ равно:

$$b_0 = \frac{1}{8}\sum_{i=1}^{8} y_i = 78{,}6;$$

значение коэффициента $b_1$ равно:

$$b_1 = \frac{1}{8}\sum_{i=1}^{8} x_{i1} y_i = \frac{-93 + 65 - 78 + 94 - 80 + 68 - 70 + 81}{8} = -1{,}62;$$

значение коэффициента $b_2$ равно:

$$b_2 = \frac{1}{8}\sum_{i=1}^{8} x_{i2} y_i = \frac{-93 + 65 + 78 + 94 - 80 - 68 + 70 + 81}{8} = 2{,}12;$$

значение коэффициента $b_3$ равно:

$$b_3 = \frac{1}{8}\sum_{i=1}^{8} x_{i3} y_i = \frac{-93 - 65 - 78 - 94 + 80 + 68 + 70 + 81}{8} = -3{,}87.$$

Следовательно, линейная модель имеет следующий вид:

$$\hat{y} = 78{,}6 - 1{,}62x_1 + 2{,}12x_2 - 3{,}87x_3.$$

Так как ковариационная матрица ПФЭ является диагональной, то полученные коэффициенты не коррелируют между собой. При исключении из модели незначимых коэффициентов нет необходимости пересчитывать остальные.

Значимость коэффициентов уравнения регрессии проверяется по критерию Стьюдента [см. формулу (5.5)]. Коэффициент $b_j$ значим, если выполняется следующее условие:

$$t_j = \frac{|b_j|}{s_{b_j}} > f_{1-\alpha,f},$$

где α – приведенный уровень значимости; $f = f_{\text{воспр}}$.

Так как диагональные коэффициенты ковариационной матрицы одинаковы и равны *N*, то оценка среднеквадратичного отклонения $s_{b_j}$ любого из коэффициентов равна:

$$s_{b_j} = \frac{s_{\text{воспр}}}{\sqrt{N}}. \qquad (5.15)$$

Для вычисления дисперсии воспроизводимости производят дублирование экспериментов в матрице планирования либо выполняют дополнительные $n_0$ опытов в центре плана. Полагаем, что в рассматриваемом примере выполнено четыре дополнительных опыта ($n_0 = 4$). По их результатам $(y_i^0: 69;\ 71;\ 73;\ 71)$ вычисляется среднее значение

$$\overline{y^0} = \frac{\sum_{i=1}^{n} y_i^0}{n_0} = 71,$$

а также дисперсия воспроизводимости

$$s^2_{\text{воспр}} = \frac{1}{n_0 - 1}\sum_{i=1}^{n}(y_i^0 - \overline{y^0})^2 = 2{,}67.$$

Следовательно, $s_{\text{воспр}} = \sqrt{2{,}67} = 1{,}63$. Тогда оценка среднеквадратичного отклонения $s_{b_j}$, согласно выражению (5.15), равна:

$$s_{b_j} = \frac{s_{\text{воспр}}}{\sqrt{N}} = \frac{1{,}63}{\sqrt{8}} = 0{,}58.$$

Критическое значение для критерия проверки: $t_{0{,}05;3} = 3{,}18$.

Получаем результаты проверки значимости коэффициентов:

$$\frac{|b_0|}{s_{b_0}} = \frac{78{,}6}{0{,}58} = 135{,}5 > t_{0{,}05;3},$$

следовательно, коэффициент $b_0$ значимый;

$$\frac{|b_1|}{s_{b_{j1}}} = \frac{|-1{,}62|}{0{,}58} = 2{,}79 < t_{0{,}05;3},$$

следовательно, коэффициент $b_1$ незначимый;

$$\frac{|b_2|}{s_{b_2}} = \frac{2{,}12}{0{,}58} = 3{,}65 > t_{0{,}05;3},$$

следовательно, коэффициент $b_2$ значимый;

$$\frac{|b_3|}{s_{b_3}} = \frac{|-3{,}87|}{0{,}58} = 6{,}67 > t_{0{,}05;3},$$

следовательно, коэффициент $b_3$ значимый.

После исключения незначимого коэффициента, уравнение регрессии принимает следующий вид:

$$\hat{y} = 78{,}6 + 2{,}12x_2 - 3{,}87x_3.$$

Для оценки адекватности модели рассчитывается величина остаточной дисперсии:

$$S^2_{\text{ост}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N - l} = \frac{\sum_{i=1}^{8}(y_i - \hat{y}_i)^2}{8 - 3} = 9{,}8,$$

где $l$ – число значимых коэффициентов регрессии.

Величина -отношения равна:

$$F = \frac{S^2_{\text{ост}}}{S^2_{\text{воспр}}} = \frac{9{,}8}{2{,}67} = 3{,}67.$$

Критическое значение критерия Фишера для $\alpha = 0{,}005$; $f_1 = f_{\text{ост}} = 5$; $f_2 = f_{\text{воспр}} = 3$ равно: $F_{\alpha,f_1,f_2} = 9$ .

Поскольку $F < F_{\alpha,f_1,f_2}$, то полученная модель адекватна.

Возвращаясь к натуральному масштабу факторов, получаем выражение регрессионной зависимости в окончательном виде:

$$\hat{y} = 78{,}6 + 2{,}12\frac{z_2 - 14}{3} - 3{,}87\frac{z_3 - 12{,}5}{5}$$

или

$$\hat{y} = 78{,}82 + 0{,}707z_2 - 0{,}774z_3.$$

Методика расчета регрессионной зависимости, содержащей эффекты от взаимодействия факторов, не отличается от рассмотренной ранее методики и не предполагает проведения дополнительных опытов.

Для расчета модели, определяемой равенством

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3$$

составляется расширенная матрица планирования, содержащая дополнительные столбцы для определения коэффициентов $b_{12}, b_{13}$ и $b_{23}$ при парных взаимодействиях (табл. 5.3):

Таблица 5.3

Расширенная матрица планирования

| Номер опыта | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_1x_2$ | $x_1x_3$ | $x_2x_3$ |
|---|---|---|---|---|---|---|---|
| 1 | +1 | –1 | –1 | –1 | +1 | +1 | +1 |
| 2 | +1 | +1 | –1 | –1 | –1 | –1 | +1 |
| 3 | +1 | –1 | +1 | –1 | –1 | +1 | –1 |
| 4 | +1 | +1 | +1 | –1 | +1 | –1 | –1 |
| 5 | +1 | –1 | –1 | +1 | +1 | –1 | –1 |
| 6 | +1 | +1 | –1 | +1 | –1 | +1 | –1 |
| 7 | +1 | –1 | +1 | +1 | –1 | –1 | +1 |
| 8 | +1 | +1 | +1 | +1 | +1 | +1 | +1 |

Коэффициенты $b_{12}, b_{13}$ и $b_{23}$ равны соответственно:

$$b_{12} = \frac{\sum_{i=1}^{n}(x_1x_2)_i\, y_i}{8} = 8{,}37,$$

$$b_{13} = \frac{\sum_{i=1}^{n}(x_1x_3)_i\, y_i}{8} = 1{,}375,$$

$$b_{23} = \frac{\sum_{i=1}^{8}(x_2x_3)_i\, y_i}{8} = -1{,}375.$$

Очевидно, что из трех коэффициентов при парных эффектах значим только коэффициент $b_{12}$. Регрессионная модель имеет следующий вид:

$$\hat{y} = 78{,}6 + 2{,}12x_2 - 3{,}87x_3 + 8{,}37x_1x_2.$$

### *5.4.2. Дробный факторный эксперимент*

Использование ПФЭ для построения модели ХТО с числом факторов большим пяти становится нецелесообразным из-за необходимости выполнения чрезмерно большого числа опытов. В этом случае правильнее будет, отказываясь от некоторых достоинств ПФЭ, перейти к выполнению ***дробного факторного эксперимента*** (ДФЭ).

Возможность сокращения числа опытов в ДФЭ по сравнению с ПФЭ рассмотрим на примере построения линейной модели ХТО с тремя факторами:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3. \qquad (5.16)$$

Очевидно, что план двухуровневого ПФЭ $2^3$ для расчета коэффициентов такой модели потребовал бы проведения восьми опытов. Сокращение числа опытов происходит за счет того, что в качестве плана ДФЭ выбирается план ПФЭ $2^2$ для меньшего числа факторов (в данном примере – для двух факторов).

Суть дробного планирования заключается в том, что дополнительно вводимый фактор $x_3$ необходимо варьировать, руководствуясь столбцом парного взаимодействия факторов $x_1$ и $x_2$.

В разделе 5.4.1 данного пособия было отмечено, что определение коэффициентов при взаимодействии факторов не требует проведения дополнительных опытов и, следовательно, для расчета всех коэффициентов рассматриваемой модели, определяемой выражением (5.16), достаточно четырех опытов.

Такой подход к построению матрицы планирования ДФЭ основан на допущении, что эффекты от взаимодействия факторов обычно существенно менее значимы по сравнению с линейными эффектами.

План, приведенный в табл. 5.4, представляет собой некоторую часть матрицы планирования ПФЭ $2^3$ – ***дробную реплику от плана ПФЭ***. Для рассматриваемого примера этот план представляет собой половину от плана $2^3$, в связи с чем используемый план ДФЭ называется ***полурепликой*** от ПФЭ $2^3$ и обозначается как $2^{3-1}$.

Таблица 5.4

Матрица планирования ДФЭ для построения трехфакторной модели

| Номер опыта | $x_0$ | $x_1$ | $x_2$ | $x_3 = x_1 x_2$ |
|---|---|---|---|---|
| 1 | +1 | +1 | +1 | +1 |
| 2 | +1 | –1 | +1 | –1 |
| 3 | +1 | +1 | –1 | –1 |
| 4 | +1 | –1 | –1 | +1 |

Использование в качестве планов ДФЭ дробных реплик планов ПФЭ позволяет сохранить для дробного планирования существенное достоинство планов ПФЭ – их ортогональность.

По мере увеличения порядка эффекта взаимодействия его влияние на функцию отклика становится менее значимым. Таким образом, влияние факторов $x_j$ на выходную переменную обычно больше, чем влияние эффекта парного взаимодействия $x_i x_j$; в свою очередь, влияние $x_i x_j$ больше, чем влияние тройного взаимодействия $x_i x_j x_s$.

Таким образом, низкая значимость межфакторных взаимодействий высокого порядка позволяет использовать их столбцы в матрице планирования для оценки линейных эффектов дополнительных факторов. Например, план ДФЭ для четырех факторов может совпадать с планом ПФЭ $2^3$ при использовании столбца тройного взаимодействия $x_1x_2x_3$ для оценки дополнительного фактора $x_4$. В этом случае план ДФЭ $2^{4-1}$ представляет собой полуреплику от плана ПФЭ $2^4$. План ПФЭ $2^3$ может быть использован и для построения модели с пятью факторами. При этом еще один вводимый фактор $x_5$ варьируется так же, как одно из парных взаимодействий: $x_1x_2$, $x_1x_3$ или $x_2x_3$. Такой план ДФЭ является ***четвертьрепликой*** от ПФЭ $2^5$ и обозначается $2^{5-2}$.

Необходимо учитывать, что достоинство ДФЭ, связанное с сокращением числа опытов, может быть сопряжено с существенным недостатком. Он обусловлен тем, что при ДФЭ уже невозможно раздельно оценивать линейные эффекты вводимых факторов и эффекты от значимых межфакторных взаимодействий. Если в рассматриваемом примере построения модели [см. формулу (5.16)] парное взаимодействие $x_1x_2$ значимо влияет на выходную переменную, то рассчитанный коэффициент модели $b_3$ является смешанной оценкой генеральных коэффициентов $\beta_3$ и $\beta_{12}$, т. к. при проведении эксперимента дополнительный фактор $x_3$ варьируется так же, как и парное взаимодействие $x_1x_2$:

$$b_3 \rightarrow \beta_3 + \beta_{12}.$$

Равенство $x_3 = x_1x_2$ называется ***генерирующим соотношением***. После умножения обеих его частей на $x_3$ и учитывая, что квадрат значения фактора на любом из двух уровней равен единице, получаем:

$$x_3^2 = x_1x_2x_3$$

или

$$1 = x_1x_2x_3.$$

Полученное выражение называется ***определяющим контрастом***. Для выяснения того, как могут быть смешаны коэффициенты, необходимо определяющий контраст умножить последовательно на $x_1$, $x_2$ и $x_3$. Получаем:

| | | |
|---|---|---|
| $x_1 = x_1^2x_2x_3;$ | $x_1 = x_2x_3;$ | $b_1 \rightarrow \beta_1 + \beta_{23};$ |
| $x_2 = x_1x_2^2x_3;$ | $x_2 = x_1x_3;$ | $b_2 \rightarrow \beta_2 + \beta_{13};$ |
| $x_1 = x_1x_2x_3^2;$ | $x_3 = x_1x_2;$ | $b_3 \rightarrow \beta_3 + \beta_{12}.$ |

При использовании для построения модели ХТО четвертьреплик или 1/8-реплик от ПФЭ, определяющих контрастов может быть несколько.

Пусть для модели ХТО с семью факторами реализован ДФЭ $2^{7-3}$, основу которого составляет матрица планирования ПФЭ $2^4$. Для оценки

влияния дополнительных факторов $x_5$, $x_6$ и $x_7$ могут быть использованы следующие генерирующие соотношения:

$$x_5 = x_1x_2x_3x_4,$$

$$x_6 = x_1x_2x_3,$$

$$x_7 = x_1x_3x_4.$$

Получаем три определяющих контраста:

$$1 = x_1x_2x_3x_4x_5,$$

$$1 = x_1x_2x_3x_6,$$

$$1 = x_1x_3x_4x_7.$$

На их основе может быть рассчитан ***полный (обобщающий) определяющий контраст***. Для этого необходимо вычислить все возможные варианты произведений частных контрастов:

$$1 = (x_1x_2x_3x_4x_5)(x_1x_2x_3x_6)(x_1x_3x_4x_7) = x_1x_3x_5x_6x_7,$$

$$1 = (x_1x_2x_3x_4x_5)(x_1x_2x_3x_6) = x_4x_5x_6,$$

$$1 = (x_1x_2x_3x_4x_5)(x_1x_3x_4x_7) = x_2x_5x_7,$$

$$1 = (x_1x_2x_3x_6)(x_1x_3x_4x_7) = x_2x_4x_6x_7.$$

Умножая полученные контрасты на переменные $x_1, x_2$, ..., получаем условия смешивания коэффициентов в рассматриваемой схеме планирования:

$$b_2 \to \beta_2 + \beta_{5,7} + \beta_{4,6,7} + \beta_{2,4,5,6} + \beta_{1,2,3,5,6,7},$$

$$b_4 \to \beta_4 + \beta_{5,6} + \beta_{2,6,7} + \beta_{2,4,6,7} + \beta_{1,3,4,5,6,7},$$

$$b_5 \to \beta_5 + \beta_{2,7} + \beta_{4,6} + \beta_{1,3,6,7} + \beta_{2,4,5,6,7}.$$

Очевидно, что линейные эффекты $x_2$, $x_4$ и $x_5$ смешаны с эффектами парного взаимодействия, что нежелательно. Поэтому необходимо рассмотреть другой вариант смешивания, например:

$$x_5 = x_1x_2x_3,$$

$$x_6 = x_1x_3x_4,$$

$$x_7 = x_2x_3x_4.$$

В этом случае получаем определяющие контрасты

$$1 = x_1x_2x_3x_5,$$

$$1 = x_1x_3x_4x_6,$$

$$1 = x_2x_3x_4x_7.$$

Обобщающий контраст:

$$1 = (x_1x_2x_3x_5)(x_1x_3x_4x_6)(x_2x_3x_4x_7) = x_3x_5x_6x_7,$$

$$1 = (x_1x_2x_3x_5)(x_1x_3x_4x_6) = x_2x_4x_5x_6,$$

$$1 = (x_1x_2x_3x_5)(x_2x_3x_4x_7) = x_1x_4x_5x_7,$$

$$1 = (x_1x_3x_4x_6)(x_2x_3x_4x_7) = x_1x_2x_6x_7.$$

Можно убедиться, что в этом варианте смешивания коэффициентов ни один линейный эффект не смешан с парными эффектами.

При использовании ДФЭ необходимо предварительно определить, какие коэффициенты модели будут представлять собой смешанные оценки генеральных коэффициентов. В зависимости от этого выбирается дробная реплика с ***наибольшей разрешающей способностью***, т. е. та реплика, которой соответствует наибольшее число линейных эффектов, не смешанных с парными.

Планы ПФЭ $2^k$ и ДФЭ $2^{k-p}$ обладают следующими достоинствами. Соответствующие им матрицы планирования ортогональны, а коэффициенты модели могут быть определены независимо друг от друга. Кроме того, при таком подходе к планированию величина определителя ковариационной матрицы $(X^TX)^{-1}$ оказывается наименьшей по сравнению с любым другим планом того же объема. Вследствие этого все коэффициенты модели определяются с равной и минимальной дисперсией $s_{b_j}^2$. Планы, обладающие такими свойствами, называются ***D-оптимальными***.

По мере удаления от центра плана точность модели убывает (возрастает величина дисперсии $s_{b_j}^2$). Если заранее не известно, в каком направлении факторного пространства будет продолжен эксперимент, то предпочтительно иметь план, точность предсказания которого одинакова в любом направлении и определяется только расстоянием от центра плана. Такое планирование называется ***ротатабельным***. Планы ПФЭ $2^k$ и ДФЭ $2^{k-p}$ являются ротатабельными.

## 5.5. Планирование эксперимента для выбора оптимального режима химико-технологического процесса

По полученной с помощью ПФЭ или ДФЭ линейной модели ХТО обычно не удается определить оптимальные значения факторов $X_j^{\text{опт}}$ $(j = 1, \dots, k)$, т. е. такие их значения, при которых функция отклика (выходная переменная) достигает экстремума. Более того, наличие адекватной линейной модели в ряде случаев оказывается признаком того, что область оптимума – ***почти стационарная область*** – еще не достигнута.

Диапазон факторного пространства, на котором функция отклика имеет почти стационарную область, показана на рис. 5.1.

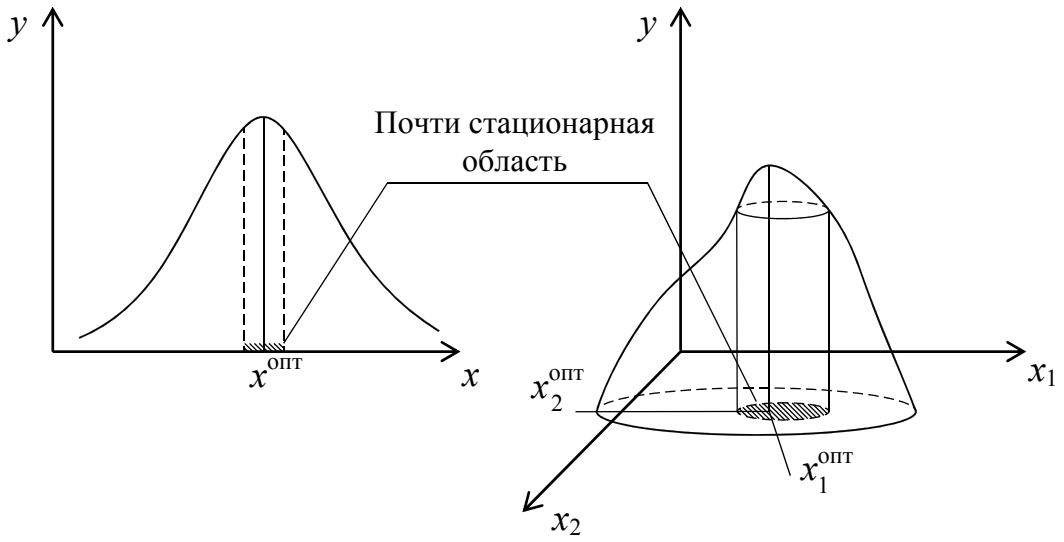


Рис. 5.1. Иллюстрация понятия «почти стационарная область» для однофакторной и двухфакторной моделей

Поэтому, используя линейную модель ХТО, осуществляют перемещение области эксперимента в факторном пространстве в направлении экстремума. Наиболее часто это реализуется с использованием метода крутого восхождения по поверхности отклика. При достижении «почти стационарной области» осуществляется ее описание с помощью регрессионной модели 2-го порядка, по которой определяются оптимальные значения факторов.

### *5.5.1. Крутое восхождение по поверхности отклика*

Самый простой способ перемещения из исходной точки факторного пространства к экстремуму (в дальнейшем будем полагать, что к максимуму) состоит в том, что поочередно, начиная с первого, изменяют каждый

из факторов до тех пор, пока значение функции отклика $y$ увеличивается. При этом остальные факторы поддерживаются на постоянных уровнях. При достижении наибольшего значения $y$ 1-й фактор фиксируют и начинают варьировать следующий фактор с целью обеспечения дальнейшего увеличения функции отклика. Описанная процедура движения к экстремуму может потребовать выполнения большого числа переборов факторов до тех пор, пока изменение любого из них не будет приводить только к уменьшению значения функции отклика.

Намного более эффективным, требующим выполнения существенно меньшего числа экспериментов, является путь к экстремуму по направлению градиента, который в каждой точке факторного пространства представляет собой направление самого крутого склона поверхности отклика, ведущего от данной точки к вершине.

Если поверхность отклика описывается некоторой функцией

$$\hat{y} = f(x_1, x_2, \dots, x_k),$$

то градиент функции равен:

$$\text{grad } f = \frac{\partial f}{\partial x_1}\vec{\imath} + \frac{\partial f}{\partial x_2}\vec{\jmath} + \cdots + \frac{\partial f}{\partial x_k}\vec{k},$$

где $\vec{\imath}, \vec{\jmath}, \vec{k}$ – единичные векторы, совпадающие с направлением координатных осей факторного пространства (орты).

Если в некоторой исходной точке факторного пространства поверхность отклика адекватно описывается линейным регрессионным уравнением

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k,$$

то $\frac{\partial f}{\partial x_1} = b_1$ , $\frac{\partial f}{\partial x_2} = b_2$, ..., $\frac{\partial f}{\partial x_k} = b_k$ и, следовательно, градиент функции отклика равен:

$$\text{grad } f = b_1\vec{\imath} + b_2\vec{\jmath} + \cdots + b_k\vec{k}.$$

То есть если изменять значения факторов пропорционально их коэффициентам $b_j$ (с учетом знака), начиная от исходной точки, то движение к оптимуму будет осуществляться по самому крутому пути. Такой процесс движения к области оптимума называется ***крутым восхождением***.

При крутом восхождении изменение каждого фактора (шаг движения) определяется как $b_j\Delta x_j$, где $\Delta x_j$ – выбранный для данного фактора интервал варьирования. Необходимо учитывать, что слишком маленький шаг потребует значительного числа опытов, а изменение функции отклика, вызванное изменением фактора, может оказаться неразличимым на фоне влияния неучтенных факторов. С другой стороны, завышение интервала варь-

ирования зачастую затрудняет получение адекватной линейной модели ХТО и, кроме того, выбор чрезмерно большого интервала варьирования может привести к «проскоку» области оптимума.

Теоретически после перемещения из исходной точки факторного пространства в новую точку необходимо поставить в ней очередную серию опытов и построить новую уточненную линейную модель ХТО. Зная значения ее коэффициента, можно скорректировать направление градиента. Однако в практических задачах для сокращения времени и средств эксперимент проводят не во всех таких восхождениях, а только в части из них.

## 5.6. Описание области факторного пространства, близкой к экстремуму. Планы 2-го порядка

Движение по графику заканчивается после достижения «почти стационарной» области. В этой области функция отклика обычно существенно нелинейная. Поэтому для описания «почти стационарной» области необходимо использовать нелинейные регрессионные модели, среди которых наибольшее распространение получили модели 2-го порядка:

$$\hat{y} = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{\substack{1 \le i \le k \\ 1 \le j \le k \\ i \ne j}} b_{ij} x_i x_j + \sum_{i=1}^{k} b_{ii} x_i^2 . \qquad (5.17)$$

Дальнейшее увеличение порядка модели приводит к существенному увеличению числа опытов. Планы проведения экспериментов для вычисления такой модели хорошо разработаны [1; 3; 4; 6]. Расчет оптимальных значений факторов по регрессии второго порядка не вызывает затруднений.

Очевидно, что для оценки адекватности модели план эксперимента должен быть ненасыщенным, т. е. число опытов в матрице планирования должно превышать число определяемых коэффициентов модели.

В -факторной модели, определяемой выражением (5.17), число коэффициентов $l$ равно:

$$l = (k+1) + C_k^2 + k = 2k + 1 + \frac{k!}{2!\,(k-2)!} =$$

$$= \frac{2(2k+1)(k-2)! + k}{2(k-2)!} = \frac{(k+1)(k+2)}{2},$$

где $C_k^2$ – число сочетаний из $k$ по 2.

Для оценки всех коэффициентов модели 2-го порядка может быть реализован ПФЭ $3^k$, при этом каждый фактор должен варьироваться на трех уровнях. Но план такого эксперимента потребует слишком большого

числа опытов, значительно превосходящего число рассчитываемых коэффициентов. Действительно, уже для модели с четырьмя факторами ($k = 4$) с числом коэффициентов $l = 28$ возникает необходимость в реализации ПФЭ $3^4$, план проведения которого состоит из 81 опыта, а для $k = 6$, когда число коэффициентов $l = 28$, план эксперимента предполагает проведение 729 опытов.

## 5.7. Центральные композиционные планы

Гораздо более рациональным является использование ***центральных композиционных планов*** (ЦКП). Композиционный план 2-го порядка формируется следующим образом: к ядру плана, представляющему собой план ПФЭ $2^k$ при числе факторов меньше пяти или к его полуреплике $2^{k-1}$, если $k \geq 5$, добавляется $n_0$ опытов в центре плана, а также $2k$ дополнительных «***звездных» точек***.

Каждому из факторов соответствуют две такие точки: при выполнении опытов в «звездных» точках плана очередной фактор принимает значение α и -α , при этом значение остальных факторов соответствуют нулевому ровню плана. Величина α называется «***звездным» плечом***. В зависимости от выбора величины «звездного» плеча и числа опытов $n_0$, ЦКП могут быть ортогональными или ротатабельными.

Общее число опытов ЦКП равно:

$$N = 2^k + 2k + n_0$$

при $k < 5$,

$$N = 2^{k-1} + 2k + n_0$$

при $k \geq 5$.

План ЦКП для двух факторов приведен в табл. 5.5. К ядру ЦКП – плану полного факторного эксперимента $2^2$ (точки 1, 2, 3, 4) – добавляют некоторое число опытов $n_0$ в центре плана (точка 5) и четыре «звездных» точки (6, 7, 8, 9) с координатами (α; 0), (–α ; 0), (0; α) и (0; –α) соответственно, где α – расстояние от центра плана до «звездной» точки («звездное» плечо).

Аналогично формируется матрица планирования ЦКП для трех факторов. Ее ядром является план ПФЭ $2^3$, к нему добавлено некоторое число точек $n_0$ в центре плана и шесть «звездных» точек с координатами (+α; 0; 0), (–α; 0; 0), (0; α; 0), (0; –α; 0), (0; 0; α) и (0; 0; –α) .

Таблица 5.5

Центральный композиционный план для построения двухфакторной модели ХТО

| Содержание плана | Номер опыта | $x_0$ | $x_1$ | $x_2$ | $x_1x_2$ | $x_1^2$ | $x_2^2$ | $y$ |
|---|---|---|---|---|---|---|---|---|
| Ядро ЦКП (ПФЭ $2^2$) | 1 | + | + | + | + | + | + | $y_1$ |
| | 2 | + | – | + | – | + | + | $y_2$ |
| | 3 | + | + | – | – | + | + | $y_3$ |
| | 4 | + | – | – | + | + | + | $y_4$ |
| Число опытов $n_0$ в центре плана | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| «Звездные» точки | 6 | + | α | 0 | 0 | $\alpha^2$ | 0 | $y_5$ |
| | 7 | + | –α | 0 | 0 | $\alpha^2$ | 0 | $y_6$ |
| | 8 | + | 0 | α | 0 | 0 | $\alpha^2$ | $y_7$ |
| | 9 | + | 0 | –α | 0 | 0 | $\alpha^2$ | $y_8$ |

### *5.7.1. Ортогональные планы 2-го порядка*

План ортогонален, если ему соответствует диагональная ковариационная матрица. Модели, полученные по ортогональным планам оценки коэффициентов, являются независимыми.

В матрице планирования ЦКП не все столбцы ортогональны, т. к.

$$\sum_{j=1}^{N} x_{0j}x_{ij}^2 \neq 0,$$

$$\sum_{j=1}^{N} x_{ij}^2 x_{hj}^2 \neq 0.$$

Так, например, в матрице ЦКП для двух факторов с числом опытов $N = 9$ имеем:

$$\sum_{j=1}^{9} x_{0j}x_{1j}^2 = \sum_{j=1}^{9} x_{0j}x_{2j}^2 = 4 + 2\alpha^2,$$

$$\sum_{j=1}^{9} x_{1j}^2 x_{2j}^2 = 4.$$

Для ортогонализации фиктивного столбца $x_0$ и столбцов $x_i^2$ необходимо преобразовать эти столбцы матрицы планирования, заменив $x_i^2$ новой переменной $x_i'$. Новая переменная $x_i'$ равна:

$$x_i^1 = x_i^2 - \frac{\sum x_{ij}^2}{N} = x_i^2 - \overline{x_i^2}.$$

После такой замены скалярные произведения столбцов $x_0$ и $x_i^1$ будут равны нулю. Имеем:

$$\sum_{j=1}^{N} x_{0j} x_{ij}^1 = \sum_{j=1}^{N} \left(x_{ij}^2 - \overline{x_i^2}\right) = \sum_{j=1}^{N} x_{ij}^2 - N\frac{\sum x_{ij}^2}{N} = 0.$$

Ортогонализация произведения столбцов $x_i^2$ и $x_h^2$ достигается выбором «звездного» плеча α.

Так, например, в матрице планирования ЦКП для двух факторов новые переменные равны:

$$x_1' = x_1^2 - \overline{x_1^2} = x_1^2 - \frac{4 + 2\alpha^2}{9},$$

$$x_2' = x_2^2 - \overline{x_2^2} = x_2^2 - \frac{4 + 2\alpha^2}{9}.$$

Поскольку на число опытов в центре плана не накладывается каких-либо ограничений, то обычно полагают $n_0 = 1$.

Величина «звездного» плеча α зависит от числа факторов. В случае двухфакторной модели $\alpha = 1$; при $k = 3$ величина «звездного» плеча $\alpha = 1{,}215$; при $k = 4$ значение $\alpha = 1{,}414$.

В табл. 5.6 приведена матрица планирования для случая $k = 2$.

Таблица 5.6

Матрица ортогонального плана 2-го порядка
для построения двухфакторной модели ХТО

| Содержание плана | Номер опыта | $x_0$ | $x_1$ | $x_2$ | $x_1x_2$ | $x_1' = x_1^2 - 2/3$ | $x_2' = x_2^2 - 2/3$ | $y$ |
|---|---|---|---|---|---|---|---|---|
| Ядро ЦКП (ПФЭ $2^k$) | 1 | 1 | + | + | + | 1/3 | 1/3 | $y_1$ |
| | 2 | 1 | − | + | − | 1/3 | 1/3 | $y_2$ |
| | 3 | 1 | + | − | − | 1/3 | 1/3 | $y_3$ |
| | 4 | 1 | − | − | + | 1/3 | 1/3 | $y_4$ |
| «Звездные» точки ($\alpha = 1$) | 5 | 1 | + | 0 | 0 | 1/3 | −2/3 | $y_5$ |
| | 6 | 1 | − | 0 | 0 | 1/3 | −2/3 | $y_6$ |
| | 7 | 1 | 0 | + | 0 | −2/3 | 1/3 | $y_7$ |
| | 8 | 1 | 0 | − | 0 | −2/3 | 1/3 | $y_8$ |
| Опыт в центре плана ($n_0 = 1$) | 9 | 1 | 0 | 0 | 0 | −2/3 | −2/3 | $y_9$ |

Благодаря ортогональности матрицы планирования, коэффициенты модели определяются независимо друг от друга по формуле

$$b_i = \frac{\sum_{j=1}^{N} x_{ij} y_j}{\sum_{j=1}^{N} x_{ij}^2}.$$

Дисперсии коэффициентов регрессии $s_{b_i}^2$ определяются по формуле

$$s_{b_i}^2 = \frac{s_{\text{воспр}}^2}{\sum_{j=1}^{N} x_{ij}^2},$$

где $s_{\text{воспр}}^2$ – дисперсия воспроизводимости.

Дисперсии коэффициентов не равны друг другу, т. к. не равны друг другу суммы элементов столбцов $(\sum_{j=1}^{N} x_{ij}^2) x_i^2$. Реализация опытов по матрице планирования с преобразованными квадратичными переменными позволяет построить модель вида

$$\hat{y} = b_0' + \sum_{1 \le i \le k} b_i x_i + \sum_{1 \le i \le l \le k} b_{il} x_i x_j + \sum_{1 \le i \le k} b_{ii} \left( x_i^2 - \overline{x_i^2} \right).$$

Для перехода к стандартной форме уравнения регрессии полагаем:

$$b_0 = b_0' - b_{11}\overline{x_1^2} - b_{22}\overline{x_2^2} - \cdots - b_{kk}\overline{x_k^2}.$$

Дисперсия этого коэффициента равна:

$$S_{b0}^2 = S_{b'}^2 + \overline{x_1^2} S_{b11}^2 + \cdots + \overline{x_{kk}^2} S_{bkk}^2.$$

Далее осуществляется проверка значимости всех коэффициентов модели *N*, проверка ее адекватности.

### *5.7.2. Ротатабельные планы 2-го порядка*

Проведение экспериментов с использованием ортогонального ЦКП обеспечивает получение регрессионной модели, коэффициенты которой не коррелированы, а значит, отпадает необходимость в пересчете коэффициентов такой модели при исключении из нее незначимых коэффициентов. Но это достоинство ортогональных ЦКП не всегда является определяющим. Решение задачи определения оптимальных значений факторов, рассмотренное в разделе 5.5.1, предполагает последовательное перемещение области проведения эксперимента в направлении к экстремуму. Если направление такого перемещения заранее не определено, то определяющим становится условие, связанное с обеспечением одинаковой точности предсказания значений функции отклика независимо от направления пе-

ремещения. Это условие выполняется при проведении экспериментов с использованием ***ротатабельных*** ЦКП. Такие планы позволяют получить регрессионную модель, обеспечивающую равенство дисперсии оценки предсказанных значений функции отклика во всех точках, равноудаленных от центра плана.

Ротатабельные ЦКП достигаются выбором величины «звездного» плеча α. Если ядром ЦКП является план ПФЭ $2^k$, то $\alpha = 2^{k/4}$, а для ядра в виде полуреплики ПФЭ величина $\alpha = 2^{\frac{k-1}{4}}$. При этом число опытов $n_0$ в центре плана не может быть произвольным, например:

1) при $k = 2$ $\alpha = 1{,}414, n_0 = 5$;
2) при $k = 3$ $\alpha = 1{,}682, n_0 = 6$;
3) при $k = 2$ $\alpha = 2, n_0 = 8$ и т. д.

Матрицы ротатабельного планирования 2-го порядка не ортогональны, поэтому объем вычисления при определении коэффициентов модели значителен.

## ЗАКЛЮЧЕНИЕ

В настоящем пособии рассмотрены вопросы построения стохастических моделей химико-технологических объектов с использованием методов корреляционного и регрессионного анализов, принципы построения оптимальных планов проведения экспериментов, обеспечивающих сокращение количества проводимых на объекте экспериментов при заданном уровне адекватности полученных моделей.

Все разделы снабжены примерами, иллюстрирующими последовательное применение алгоритмов, используемых при построении стохастических моделей от этапа предварительной обработки экспериментальных данных до проверки адекватности полученной регрессионной зависимости.

## БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Ахназарова, С. Л. Методы оптимизации эксперимента в химической технологии : учеб. пособие для вузов / С. Л. Ахназарова, В. В. Кафаров. – М. : Высш. шк., 1985. – 319 с.

2. Гнеденко, Б. В. Курс теории вероятностей : учеб. для вузов / Б. В. Гнеденко. – М. : Наука, 1988. – 448 с.

3. Бондарь, А. Г. Планирование эксперимента при оптимизации процессов химической технологии (алгоритмы и примеры) / А. Г. Бондарь, Г. А. Статюха, И. А. Потяженко. – Киев : Вища шк., 1980. – 264 с.

4. Саутин, С. Н. Планирование эксперимента в химии и химической технологии / С. Н. Саутин. – Л. : Химия, 1975. – 48 с.

5. Львовский, Е. Н. Статистические методы построения эмпирических функций : учеб. пособие для студентов втузов / Е. Н. Львовский. – М. : Высш. шк., 1988. – 239 с.

6. Теория планирования эксперимента [Электронный ресурс] / Режим доступа: http://appmath.narod.ru.

7. Формулы математической статистики [Электронный ресурс] / Режим доступа: http://www.matburo.ru/ms_spr.php.

8. Наливайко, Т. Е. Теоретическое обоснование системы критериев и показателей сформированности компетенстностей обучающихся / Т. Е. Наливайко, М. В. Шинкорук // Ученые записки Комсомольского-на-Амуре гос. техн. ун-та. Науки о человеке, обществе и культуре. – 2013. – № I-2(13). – С. 23-30.

*Учебное издание*

**Гринфельд** Григорий Михайлович
**Моисеев** Андрей Владимирович

# МЕТОДЫ ОПТИМИЗАЦИИ ЭКСПЕРИМЕНТА В ХИМИЧЕСКОЙ ТЕХНОЛОГИИ

Конспект лекций

Научный редактор – доктор технических наук,
профессор В. В. Петров

Редактор Е. В. Назаренко

Подписано в печать 30.05.2014.
Формат 60 × 84 1/16. Бумага 60 г/м$^2$. Ризограф EZ570e.
Усл. печ. л. 4,65. Уч.-изд. л. 4,32. Тираж 50 экз. Заказ 26294.

Редакционно-издательский отдел
Федерального государственного бюджетного образовательного учреждения
высшего профессионального образования
«Комсомольский-на-Амуре государственный технический университет»
681013, Комсомольск-на-Амуре, пр. Ленина, 27.

Полиграфическая лаборатория
Федерального государственного бюджетного образовательного учреждения
высшего профессионального образования
«Комсомольский-на-Амуре государственный технический университет»
681013, Комсомольск-на-Амуре, пр. Ленина, 27.