

Парадигма развития науки
Методологическое обеспечение

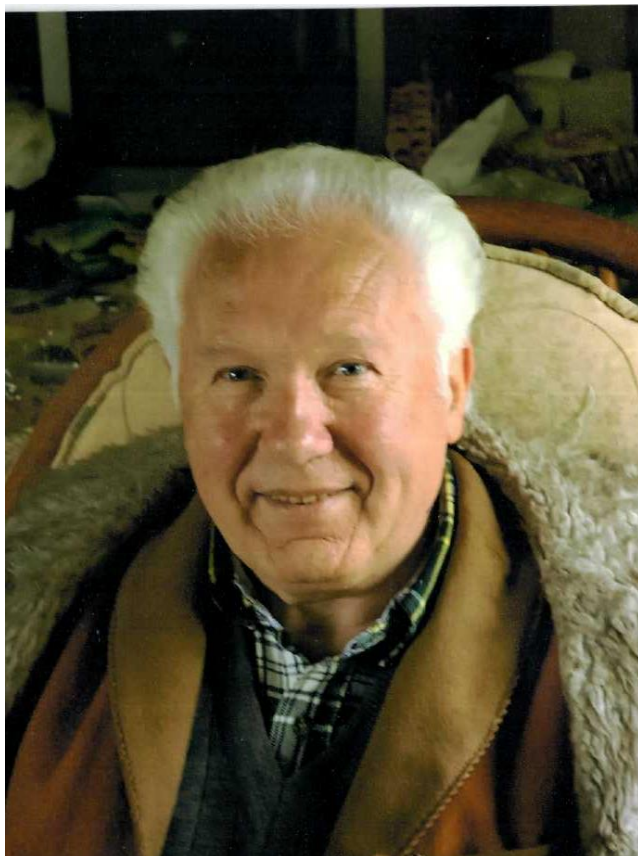
А.Е. Кононюк

**ОСНОВЫ НАУЧНЫХ
ИССЛЕДОВАНИЙ**

(Общая теория эксперимента)

Книга 1

**Киев
КНТ
2010**



Кононюк Анатолий Ефимович



УДК 519. 237.5:515.126.2

ББК 22.172+22.152

К 15

Рецензент:

Н.К.Печурин - д-р техн. наук, проф. (Национальный авиационный университет).

Кононюк А.Е.

К15 Основы научных исследований (общая теория эксперимента)

- В 4-х кн.-К.1. -К.: 2011.- 508 с.

ISBN 978-966-373-696-9 (многотомное издание)

ISBN 978-966-373- (книга 1)

Серия книг посвящена основам научного исследования, методы которых широко используются в экспериментальных исследованиях. Подробно рассмотрена техника факторного планирования эксперимента. Особое внимание уделено вопросам интерпретации результатов эксперимента и проверки правильности исходных предпосылок, приемам планирования в лабораторных и промышленных условиях, в том числе блочному планированию. Изложение сопровождается детальным анализом большого числа практических примеров.

Монография рассчитана на магистров, аспирантов, докторантов и других научных работников различного профиля.

УДК 519. 237.5:515.126.2

ББК 22.172+22.152

ISBN 978-966-373-696-9 (многотомное издание)

ISBN 978-966-373- (книга 1)

© Кононюк А.Е., 2011

© КТН, 2011

Оглавление

Предисловие	7
Введение	9
1. Творчество и научные исследования	13
1.1. Основные понятия и определения.....	13
1.2. О научном исследовании.....	28
1.3. Методология теоретических и экспериментальных исследований.....	38
1.3.1. Законы и формы мышления (мышление, понятие, абстракция).....	38
1.3.2. Законы и формы мышления (сравнение, индукция и дедукция, анализ и синтез).....	40
1.3.3. Законы и формы мышления (обобщение, аналогия, гипотеза).....	43
1.3.4. Методология исследований.....	45
1.3.5. Задачи теоретических исследований.....	46
1.4. Цель и предмет исследования.....	47
1.4.1. Цель исследования.....	47
1.4.2. Предмет исследования.....	52
1.5. Рабочая гипотеза.....	54
1.6. Особенности исследования непрерывных процессов.....	61
1.6.1. Постановка задач исследования сложных процессов.....	62
1.6.2. Основные цели и задачи процесса исследования технологических систем.....	67
1.7. Введение в методы исследования.....	87
1.7.1. Измерения и отметки.....	88
1.7.2. Точность измерений.....	95
1.7.3. Наблюдение.....	111
1.7.4. Поисковые опыты.....	112
1.7.5. Основные опыты.....	116
1.7.6. Определение общего количества опытов.....	122
2. Погрешности измерений и устройств	135
2.1. Погрешности измерений.....	135
2.1.1. Измеряемые сигналы.....	135
2.1.2. Погрешности измерений.....	141
2.1.3. Случайные погрешности.....	146
2.1.4. Числовые характеристики.....	149
2.1.5. Нормальный закон распределения.....	153
2.1.6. Закон равномерной плотности распределения.....	155
2.1.7. Композиции законов распределения. Правила суммирования случайных погрешностей.....	158

2.1.8. Зависимые и независимые случайные величины.	
Корреляция случайных погрешностей.....	163
2.1.9. Определение законов распределения случайных величин	167
2.1.10. Числовые характеристики статистического распределения.....	171
2.1.11. Проверка правдоподобия гипотезы о соответствии статистического закона теоретическому.....	180
2.1.12. Погрешности косвенного измерения.....	183
2.1.13. Промахи.....	186
2.1.14. Систематические погрешности. Оценка неисключенного остатка.....	187
2.1.15. Запись результата измерения.....	189
2.2. Погрешности измерительных устройств при неизменном во времени значении измеряемой величины.....	191
2.2.1. Структурные схемы измерительных устройств.....	191
2.2.2. Характеристики преобразователей.....	193
2.2.3. Погрешности преобразователей. Аддитивная и мультипликативная составляющие.....	197
2.2.4. Частные погрешности. Причины их возникновения и законы распределения.....	202
2.2.5. Методические погрешности.....	213
2.2.6. Суммирование частных погрешностей.....	217
2.2.7. Нормирование погрешностей измерительных устройств.....	223
2.3. Погрешности измерительных устройств при изменении значения измеряемой величины во времени.....	226
2.3.1. Понятие динамической погрешности.....	226
2.3.2. Представление измеряемых сигналов стандартизованными воздействиями.....	229
2.3.3. Реакция измерительных устройств на синусоидально изменяющийся входной сигнал.....	236
2.3.4. Реакция измерительных устройств на единичную функцию и δ -функцию.....	252
2.3.5. Связь между переходными, импульсными и частотными характеристиками измерительных устройств.....	258
2.3.6. Реакция измерительного устройства на линейно-возрастающую во времени функцию.....	259
2.3.7. Оценка точности воспроизведения измерительными устройствами изменяющихся во времени сигналов.....	262
2.3.8. Случайные сигналы и их характеристики.....	268
2.3.9. Реакция измерительных устройств на случайные возмущения. Характеристики точности.....	274

2.3.10. Определение динамических свойств измерительных устройств с помощью моделей-аналогов	277
3. Параметры, факторы, модели объектов исследования	284
3.1. Параметры оптимизации	284
3.1.1. Виды параметров оптимизации	285
3.1.2. Требования к параметру оптимизации	288
3.1.3. Задачи с несколькими выходными параметрами	295
3.2. Обобщенный параметр оптимизации	298
3.2.1. Способы построения обобщенного отклика	298
3.2.2. Шкала желательности (предпочтительности)	302
3.2.3. Преобразование частных откликов в частные функции желательности	306
3.2.4. Обобщенная функция желательности	309
3.3. Факторы	313
3.3.1. Определение фактора	314
3.3.2. Требования, предъявляемые к факторам при планировании эксперимента	317
3.3.3. Требования к совокупности факторов	320
3.4. Выбор модели	324
3.4.1. Модели объектов исследования	324
3.4.2. Шаговый принцип	332
3.4.3. Полиномиальные модели	336
4. Основные методы планирования экстремального эксперимента	344
4.1. Сущность теории планирования эксперимента	344
4.2. Принятие решений перед планированием эксперимента	348
4.3. Корреляционный анализ	361
4.3.1. Понятие о корреляции	361
4.3.2. Линейная корреляция	365
4.3.3. Нелинейная корреляция	383
4.3.4. Множественная линейная корреляция	384
4.3.5. Корреляционные (регрессионные) уравнения	387
4.3.6. Ранговая корреляция	397
4.4. Дисперсионный анализ	401
4.4.1. Однофакторный дисперсионный анализ	401
4.4.2. Двухфакторный дисперсионный анализ	406
4.4.3. Дисперсионный анализа при использовании схемы латинского квадрата	411
4.5. Регрессионный анализ	416
4.5.1. Метод наименьших квадратов (МНК)	418
4.5.2. Статистический анализ уравнения регрессии	426

4.5.3. Матричный подход к регрессионному анализу.....	434
4.5.3.1. Метод наименьших квадратов для одного фактора.....	434
4.5.3.2. Некоторые операции над матрицами.....	436
4.5.3.3. Обобщение метода наименьших квадратов на многофакторный линейный случай.....	442
4.5.3.4. Статистический анализ.....	447
4.5.3.5. Взвешенный метод наименьших квадратов и статистический анализ.....	453
4.5.3.6. Критерии оптимальности планов.....	466
4.5.4. Построение регрессионной модели.....	472
4.5.4.1. Выбор модели.....	473
4.5.4.2. Эксперимент.....	474
4.5.4.3. Вычисление оценок коэффициентов модели.....	476
4.5.4.4. Проверка модели на адекватность.....	482
4.5.4.5. Упрощение модели.....	484
4.5.4.6. Точность регрессионной модели.....	487
Приложения	494
Список литературы	507

Предисловие

Способы наилучшей организации эксперимента, обработки и интерпретации его результатов привлекают внимание как математиков, так и прикладников-экспериментаторов.

Интерес к науке об эксперименте не случаен. Он вызван широкими масштабами экспериментальных исследований и обусловлен значительным экономическим эффектом, к которому приводит оптимальная организация эксперимента.

Любой эксперимент включает следующие основные этапы:

- а) постановка задачи и выбор плана эксперимента;
- б) проведение эксперимента, сбор и обработка данных;
- в) анализ результатов эксперимента.

Чрезвычайно ответственным является этап анализа результатов, ибо именно здесь делаются выводы и принимаются решения по совершенствованию исследуемой системы.

Чтобы выводы были надежными, следует всесторонне исследовать исходные данные, проверяя различные предположения о причинах получения тех или иных результатов. Поверхностный подход к анализу данных и формальное использование математического аппарата могут привести к существенным просчетам при принятии решений. Именно на эти аспекты обращается внимание в данной работе, где подробно разбирается большое число примеров интерпретации результатов конкретных экспериментов, а также приводится множество полезных практических приемов и рекомендаций, относящихся к анализу данных, проверке исходных предпосылок, выбору планов эксперимента.

Методы теории экспериментов позволяют во многих случаях уменьшить число опытов и таким образом достичь заданных целей с экономией значительных ресурсов (времени, средств на проведение экспериментов). Однако больший эффект можно получить, если весь логический путь исследования реализовать на основе системы методов теории эксперимента. В этом случае можно не только сэкономить ресурсы, но и получить новые качественные результаты, подготовить необходимую информацию для проектирования объектов.

Успех экспериментирования в значительной мере обязан теории эксперимента, которая признана дать экспериментатору ответы на следующие вопросы:

1. Как нужно организовать эксперимент, чтобы наилучшим образом решить поставленную задачу (в смысле затрат времени, средств или точности результатов)?

2. Как следует обрабатывать результаты эксперимента, чтобы получить максимальное количество информации об исследуемом объекте?

3. Какие обоснованные выводы можно сделать об исследуемом объекте по результатам эксперимента?

Цель данной работы — расширить и углубить идею использования системного подхода при производстве эксперимента. Для этого предполагается выделить и рассмотреть основные методы планирования эксперимента, применяющиеся для оптимизации и получения математических моделей объектов, представить их алгоритмами, показать их работоспособность при решении практических задач, выделить основные цели процесса исследования и указать место алгоритмов в логической схеме исследования.

Для эффективного усвоения изложенного в работе материала пользователю достаточно знаний в области теории вероятностей, математической статистики и матричной алгебры, предусмотренных вузовским курсом математики. Для облегчения поиска вспомогательной информации в приложении к пособию приведены некоторые положения математической статистики, теории случайных функций и таблицы, необходимые для решения примеров и задач.

Введение

На современном этапе научно-технического прогресса необыкновенно возрастает роль средств, позволяющих рационально использовать ресурсы, выделенные для решения научно-технических задач. Кибернетика предлагает такие средства, как исследование операций, теория систем, математическое моделирование, теория эксперимента, вычислительная техника и др.

Часть этих методов предназначена для увеличения эффективности научного эксперимента на всех стадиях разработки, исследования, проектирования и эксплуатации объектов. Единство теории и практики эксперимента совместно с вычислительной техникой образуют комплекс автоматизированного эксперимента, предназначенный для повышения производительности научного труда.

Объекты, на которых проводятся эксперименты, отличаются прежде всего протекающими в них процессами.

Объект, на котором осуществляется планируемый эксперимент, характеризуется обязательным условием — все входные переменные, или факторы, x_1, x_2, \dots, x_n должны быть управляемыми. Этого требует сама постановка планирования эксперимента, предполагающего активное вмешательство в ход эксперимента. Такой объект называют объектом исследования.

Согласно принятой терминологии

- *факторы* x_1, x_2, \dots, x_n — это измеряемые и регулируемые входные переменные объекта (независимые переменные);

- *помехи* z_1, z_2, \dots, z_s — это неконтролируемые, случайным образом изменяющиеся переменные объекта;

- *выходные параметры* y_1, y_2, \dots, y_m — это контролируемые переменные, которые определяются факторами и связаны с целью исследования.

В планируемом эксперименте y называют параметром оптимизации (технологический или экономический показатель процесса).

Факторы x_1, x_2, \dots, x_n иногда называют основными, поскольку они определяют условия эксперимента. Помехи z_1, z_2, \dots, z_s — как правило, недоступны для измерения. Они проявляются лишь в том, что изменяют влияние факторов на выходные переменные. Объект исследования может иметь несколько выходных переменных. Опыт показывает, что в большинстве случаев удается ограничиться одним параметром оптимизации, и тогда вектор Y превращается в скаляр y .

Количество факторов и характер их взаимосвязей с выходной переменной определяют сложность объекта исследования. При наличии качественной статистической информации о факторах и

зависящей от них выходной переменной можно построить математическую модель объекта исследования или функцию отклика $y=f(x_1, x_2, \dots, x_n)$, связывающую параметр оптимизации с факторами, которые варьируются при проведении опытов.

Пространство с координатами x_1, x_2, \dots, x_n принято называть факторным, а графическое изображение функции отклика в факторном пространстве — поверхностью отклика.

При описании объектов, находящихся в стационарном состоянии, математическая модель чаще всего представляется полиномом:

$$y = f(x_1, x_2, \dots, x_n, \beta_0, \beta_1, \dots, \beta_n). \quad (B.1)$$

Поскольку в реальном процессе всегда существуют неуправляемые и неконтролируемые переменные, величина y носит случайный характер. Поэтому при обработке экспериментальных данных получаются так называемые выборочные коэффициенты регрессии $b_0, b_1, \dots, b_i, \dots, b_n$, являющиеся оценками теоретических коэффициентов

$$\beta_0, \beta_1, \dots, \beta_i, \dots, \beta_n.$$

Тогда математическая модель в форме уравнения регрессии в общем случае будет иметь вид

$$\hat{y} = b_0 + \sum_i b_i x_i + \sum_{i < j} b_{ij} x_i x_j + \sum_i b_{ii} x_i^2 + \dots \quad (B.2)$$

Если анализируются нестационарные, т. е. изменяющиеся во времени состояния объекта, приходится рассматривать не случайные величины, как ранее, а случайные процессы. Случайный процесс можно рассматривать как систему, состоящую из бесконечного множества случайных величин. При моделировании таких объектов использовать модель в виде (B.2) уже недопустимо — необходимо переходить к специальным интегрально-дифференциальным моделям.

В отличие от статистических математических моделей существуют физико-химические модели (иногда их называют аналитическими). При построении таких моделей используются физико-химические закономерности моделируемых процессов. Построение физико-химических моделей сопряжено с проведением обширных и длительных исследований, поскольку необходимо рассматривать микроструктуру процессов (чаще это иерархия процессов) и описывать их математически. Модели в этом случае представляются системами алгебраических и различных дифференциальных уравнений. Не всегда удается рассчитать численные значения коэффициентов, входящих в эти системы, и тогда опять прибегают к эксперименту. Используя теорию случайных процессов, удастся найти численные значения

коэффициентов модели по данным статистической информации. Эта задача известна в теории как *задача идентификации*.

Составлению плана эксперимента всегда должны предшествовать сбор априорной информации для составления характеристики объекта исследования, опыты по наладке экспериментальной установки и при необходимости — опыты для установления области определения наиболее существенных факторов и выходной переменной.

Теорией и практикой эксперимента выработаны определенные требования (условия), которым должны удовлетворять независимые и зависимые переменные. Поэтому на стадии подготовки к проведению эксперимента весьма полезны приведенные ниже рекомендации.

1. При выборе *выходной переменной* необходимо учитывать, что она должна: иметь количественную характеристику, т. е. должна измеряться; однозначно оценивать (измерять) работоспособность объекта исследования; быть статистически эффективной, т. е. иметь возможно меньшую дисперсию при проведении опытов (это позволяет четко различать опыты); отражать как можно более широкий спектр исследуемого явления, т. е. обладать универсальностью (практически это требование обеспечить трудно, тогда рекомендуют пользоваться так называемой обобщенной переменной); иметь достаточно четкий физический смысл.

Удачный выбор выходной переменной в значительной степени определяется уровнем знания объекта исследования.

2. При выборе *факторов* нужно выполнять следующие требования: фактор должен быть регулируемым, т.е. определенным регулирующим устройством фактор должен изменяться от значения x'_i до значения x''_i ; (например, расход вещества может изменяться от 30 до 40 л/ч или количество вещества A в смеси от 10 до 80%, это свойство называют *операционной определенностью*); точность измерения и управления фактором должна быть известна и достаточно высока (хотя бы на порядок выше точности измерения выходной переменной), очевидно, что низкая точность измерения фактора уменьшает возможности воспроизведения эксперимента; связь между факторами должна быть как можно меньшей (в пределе должна отсутствовать), это свойство называют *однозначностью факторов*, что соответствует независимости одного фактора от другого.

Ряд требований предъявляется одновременно к *факторам* и *выходной переменной*: факторы и выходная переменная должны иметь области определения, заданные технологическими или принципиальными ограничениями (пример технологического ограничения — максимальная производительность компрессора,

подающего газ в реактор; пример принципиального — температура кристаллизации жидкого продукта, образующегося в результате реакции); области определения факторов должны быть таковы, чтобы при их предельных значениях значение выходной переменной оставалось в своих границах; между факторами и выходной переменной должно существовать однозначное соответствие (причинно-следственная связь).

Успех современного экспериментирования в значительной мере обязан теории эксперимента, которая признана дать экспериментатору ответы на следующие вопросы:

1. Как нужно организовать эксперимент, чтобы наилучшим образом решить поставленную задачу (в смысле затрат времени, средств или точности результатов)?
2. Как следует обрабатывать результаты эксперимента, чтобы получить максимальное количество информации об исследуемом объекте?
3. Какие обоснованные выводы можно сделать об исследуемом объекте по результатам эксперимента?

Как уже упоминалось, основой теории эксперимента является статистическое представление об эксперименте (рассматриваются случайные величины или случайные функции). Это представление отвечает действительности; как правило, итоги эксперимента связаны с некоторой неопределенностью, получающейся в результате влияния неконтролируемых факторов, случайного характера процесса на микроуровне, изменений условий эксперимента, ошибок измерения и др.

Статистическое представление об эксперименте позволило перейти к многофакторному активному эксперименту, в котором удается надежно отделить влияние факторов от шумового фона и перейти к статистически обоснованным методам анализа результатов и принятия решений.

Теория эксперимента использует принципы и концепции кибернетики, такие как математическое моделирование, оптимизация. Большое значение имеет концепция оптимального использования факторного пространства с точки зрения определенных характеристик точности решения. Так, удается получать оценки коэффициентов модели, обладающие минимальной дисперсией, или удается получать наилучшие предсказательные свойства модели.

Теория эксперимента указывает исследователю точную логическую схему и способы поиска решения задач на разных этапах исследования. Можно представить весь процесс исследования циклами,

повторяющимися после решения каждой из последовательных задач исследования, причем объем знаний об объекте непрерывно увеличивается.

Цель настоящей работы состоит в описании эффективных методов эксперимента, широко используемых при решении задач лабораторных и промышленных исследований.

1. Творчество и научные исследования

1.1. Основные понятия и определения

При изложении теории основ научных исследований используются понятия, относящиеся к общей теории систем. В общей теории систем такие понятия в силу их общности трактуются слишком широко, поэтому, чтобы достичь однозначности в понимании последующего материала, определим эти понятия применительно к основам научных исследований

Нам предстоит давать определения и разъяснять специальные термины на протяжении всей настоящей работы, но некоторые наиболее общие термины приводятся ниже, ибо без них мы не сможем понимать суть излагаемого материала.

Элемент — это минимальный неделимый объект. Свойства элемента— это его свойства как целого. Элемент можно использовать только как целое, поэтому недопустимо говорить о половине или четверти элемента. Неделимость элемента — это прежде всего удобное понятие, но не физическое свойство. Опирируя понятием «элемент», исследователь оставляет за собой право перейти на другой уровень рассмотрения вопросов и говорить о том, из чего состоит элемент, а это свидетельствует о физической разложимости последнего. Таким образом, объекты называются элементами по соглашению, принимаемому с целью дать ответ на определенные вопросы, стоящие перед исследователями. Изменение вопросов может потребовать разложения элементов на составные части или объединения нескольких элементов в один.

Система — это совокупность связанных элементов, объединенных в одно целое для достижения определенной цели.

Здесь под **целью** понимается совокупность результатов, определяемых назначением системы.

Наличие цели и заставляет связывать элементы в систему.

Целостность — наиболее важное свойство системы. Элемент принадлежит системе потому, что он связан с другими ее элементами, так что **множество элементов, составляющих систему,**

невозможно разбить на два и более несвязанных подмножества. Удаление из системы элемента или совокупности элементов непременно изменяет ее свойства в направлении, отличном от цели.

Искусственные (инженерные) системы описывают путем определения их **функций и структур.**

Функция системы — это правило(а) получения результатов, предписанных целью (назначением) системы.

Определяя функцию системы, ее поведение описывают с использованием некоторой системы понятий — отношений между переменными, векторами, множествами и т. п. Функция устанавливает, что делает система для достижения поставленной цели безотносительно к физическим средствам, (элементам, связям), составляющим саму систему, и не определяет, как устроена система. Системы изучают на разных уровнях абстракции, с использованием различных подходов, каждый из которых дает ответ на определенные вопросы. В связи с этим функции системы могут описываться с разной степенью детализации. Для описания функций систем используются теории множеств, алгоритмов, случайных процессов, и т. д.

Функционировать — значит реализовать функцию, т. е. получать результаты, предписанные назначением системы.

Структура системы — это фиксированная совокупность элементов и связей между ними. В общей теории систем под структурой принято понимать **только множество связей между элементами**, т. е. структура понимается как картина, **отображающая только конфигурацию системы безотносительно к составляющим ее элементам.** Такое толкование ее понятия удобно при структурном подходе к изучению свойств различных систем — систем с параллельными, последовательными, иерархическими структурами, обратными связями и т. п. На практике в понятие «структура» включают не только множество связей, но и множество элементов, между которыми существуют связи. Этот смысл отражен в данном определении структуры. Наиболее часто структура системы изображается в форме графа: элементы системы представляются вершинами графа, а связи — дугами (ребрами) графа.

Граф — это математическая форма отображения структур.

Инженерной формой изображения структур систем являются **схемы.** Схема и граф — понятия адекватные **по содержанию**, но различные **по форме.** В схемах элементы и связи обозначаются любыми фигурами, удобными для инженерных (производственных) применений.

Организация — это явление аранжировки с целью осуществления определенных функций в системах, состоящих из большого числа элементов.

Обычно к одной и той же цели можно прийти различными способами, исходя из различных принципов организации систем. Каждый принцип организации задает определенный способ построения множества систем, аналогичных по назначению, но различных по функциям и структурам. Конкретная система представляет собой лишь пример реализации некоторого способа организации. Например, подавляющее большинство современных ЭВМ строится на основе одного принципа организации — принципа программного управления реализацией алгоритма на основе команд, имеющих операционно-адресную структуру. Таким образом, **организация** — понятие более высокого ранга, чем функция и структура; **организация** — это модель, на основе которой могут строиться многие конкретные системы.

Если речь идет о способе порождения функций, достаточных для достижения определенной цели (определенных результатов), то используется термин **функциональная организация**. Если же речь идет о наборе элементов и способе их соединения в структуру, обеспечивающую реализацию функций определенного класса, то используется термин - **структурная организация**. Определяя некоторый способ функциональной организации, выявляют класс функций, присущих системам определенного назначения (безотносительно к средствам, необходимым для реализации этих функций), а определяя способ структурной организации, выявляют правило построения структур, реализующих некоторый класс функций, т. е. отвечающих некоторому назначению.

Анализ — это процесс определения свойств, присущих системе. Типичная задача анализа состоит в следующем. Известны функции и характеристики элементов, входящих в состав системы, и определена структура системы. Необходимо определить функции или характеристики, присущие совокупности элементов в целом.

Синтез — это процесс порождения функций и структур, необходимых и достаточных для получения определенных результатов. Выявляя функции, реализуемые системой, определяют некоторую абстрактную систему, о которой известно только то, что она будет делать. В связи с этим этап синтеза функций называется **абстрактным синтезом**, а этап порождения структуры, реализующей заданные функции, — **структурным синтезом**.

Эффективность — это степень соответствия системы своему назначению. Из двух систем более эффективной считается та, которая лучше соответствует своему назначению. Оценка эффективности системы — одна из задач анализа систем.

Показатель эффективности (качества)—это мера одного свойства (характеристики) системы. Показатель эффективности всегда имеет **количественный** смысл, т. е. является измерением некоторого свойства. По этой причине использование некоторого показателя эффективности предполагает наличие способа измерения (оценки) значения этого показателя. Для оценок эффективности систем могут применяться, например, такие показатели, как производительность, стоимость, надежность, габариты и т. п.

Критерий эффективности — это мера эффективности системы. Критерий эффективности имеет количественный смысл и измеряет степень эффективности системы, обобщая все ее свойства в одной оценке — значении критерия эффективности. Эффективность систем, создаваемых для одной цели, оценивается на основе одного критерия, общего для этого класса систем. Различие в назначениях систем предполагает, что для оценки эффективности таких систем используются различные критерии. Если при увеличении эффективности значение критерия возрастает, то критерий называется **прямым**; если значение критерия уменьшается, то **инверсным**. Из двух систем более эффективной считается та, которой соответствует большее значение прямого критерия (меньшее значение инверсного критерия).

Оптимальная система — это система, которой соответствует максимальное (минимальное) значение прямого (инверсного) критерия эффективности на множестве мыслимых вариантов построения систем.

Модель — это физическая или абстрактная система, адекватно представляющая собой объект исследования.

Физические модели образуются из совокупности материальных объектов. Для их построения используются различные свойства, причем природа применяемых в модели материальных элементов не обязательно та же, что и в исследуемом объекте. Примером физической модели является макет.

Абстрактная модель — это описание объекта исследований на некотором языке. Абстрактность модели проявляется в том, что **ее компонентами являются понятия, а не физические элементы** (например, словесные описания, чертежи, схемы, графики, таблицы, программы (алгоритмы), математические описания).

Необходимое условие для перехода от исследования объекта к исследованию модели и дальнейшего перенесения его результатов на объект исследования — требование адекватности модели и объекта.

Адекватность предполагает воспроизведение моделью с необходимой полнотой всех свойств объекта, существенных для целей данного исследования.

Понятие адекватности — весьма широкое и основывается на строгих в математическом отношении понятиях **изоморфизма и гомоморфизма**.

Две системы, в данном случае объект исследования и его модель, называются **изоморфными**, если между ними существует такое взаимно-однозначное соответствие, что соответствующие объекты различных систем обладают соответствующими свойствами и находятся в соответствующих отношениях друг с другом.

В общем случае обеспечение изоморфизма модели и объекта исследования может быть не только трудновыполнимым, но и излишним, поскольку сложность модели при этом может оказаться настолько значительной, что никакого упрощения исследовательской задачи не произойдет.

Гомоморфизм, так же как и изоморфизм, предполагает сохранение в модели всех определенных на объекте исследования свойств и отношений. Однако требование взаимно-однозначного соответствия заменяется требованием однозначного соответствия модели объекту, тогда как обратное соответствие — соответствие объекта модели — не однозначно.

Концептуальная модель — это абстрактная модель, выявляющая причинно-следственные связи, присущие исследуемому объекту и существенные в рамках определенного исследования. **Основное назначение концептуальной модели** — выявление набора причинно-следственных связей, учет которых необходим для получения требуемых результатов. Один и тот же объект может представляться различными концептуальными моделями, которые строятся в зависимости от цели исследования. Так, одна концептуальная модель может отображать временные аспекты функционирования системы, другая — влияние отказов на работоспособность системы и т. д.

Математическая модель — это абстрактная модель, представленная на языке математических отношений. Математическая модель имеет форму функциональных зависимостей между параметрами, учитываемыми соответствующей концептуальной моделью. Эти зависимости конкретизируют причинно-следственные связи,

выявленные в концептуальной модели, и характеризуют их количественно.

Моделирование — это процесс представления объекта исследований адекватной ему моделью и проведения экспериментов с моделью с целью получения информации об объекте исследования.

При моделировании модель выступает и как средство, и как объект исследований.

Моделирование является косвенным методом выявления свойств объекта в том смысле, что исследование производится не над самим объектом, а над представляющей объект системой — моделью.

Отличительная особенность моделирования как метода исследования заключается в возможности исследования таких объектов, прямой эксперимент с которыми трудновыполним, экономически невыгоден или вообще невозможен.

Одним из основополагающих терминов в теории научных исследований является понятие «**эксперимент**». Однако в научной и справочной литературе нет единого определения этого термина. Поэтому ниже будет приведено несколько определений понятия «эксперимент», заимствованных из изданий главным образом справочного характера, где делалась попытка обобщить имеющиеся по этому поводу высказывания.

1. *Эксперимент* (от латинского *experimentum*—проба, опыт) — чувственно-предметная деятельность в науке, осуществляемая теоретически познанными средствами. В научном языке термин «эксперимент» обычно используется интуитивно в значении, общем для целого ряда сопряженных понятий: опыт, целенаправленное наблюдение, воспроизведение объекта познания, организация особых условий его существования, проверка предсказания и т. п. (Философская энциклопедия. Т. 5. М., «Советская энциклопедия», 1970).

2. *Эксперимент* (латинский *experimentum* — проба, опыт) — 1) научно поставленный опыт, наблюдение исследуемого явления в точно учитываемых условиях, позволяющих следить за ходом явлений и воссоздать его при повторении этих условий; 2) опыт, попытка (Словарь иностранных слов. М., «Советская энциклопедия», 1964).

3. *Эксперимент* — систематическое изменение условий необходимого явления и связи его с другими с целью выяснения его природы, его происхождения и методов сознательного овладения данным процессом. Кювье так определяет задачи эксперимента: «Наблюдатель слушает природу, экспериментатор вопрошает и принуждает ее разоблачиться» (БСЭ, 1-е изд., т. 63, 1933).

4. *Эксперимент* — научно поставленный опыт, наблюдение исследуемого явления в точно учитываемых условиях, позволяющих следить за ходом явления и воссоздать его каждый раз при повторении этих условий (БСЭ, 2-е изд., т. 48, 1957).

5. *Эксперимент* — операция, предназначенная для обнаружения истины, принципа или эффекта или, после их обнаружения, для уточнения или иллюстрации. Он отличается от наблюдения тем, что наблюдаемые явления в большей или меньшей степени контролируются человеком (Encyclopedia Americana, v. 10, 1944).

6. *Эксперимент*: 1) испытывать или подвергать испытанию, испытание, проверка; 2) средство или лекарство, предназначенное для испытания; 3) действие или операция, предпринятые с целью обнаружения нового или проверки гипотезы, или иллюстрации известной истины; 4) подробная процедура, метод, система явлений или последовательность действий, принятые в состоянии неуверенности относительно того, отвечают ли они цели (Oxford English Dictionary, 1958).

7. Само по себе понятие «эксперимент» в физике означает действие, направленное на искусственное создание условий для осуществления того или иного физического явления и для наблюдения этого явления в условиях, по возможности наиболее чистых, т. е. не осложняемых другими физическими явлениями (Жданов Г, В, —В кн.: Современный детерминизм. Законы природы, М., «Мысль», 1973).

Как видно, что даже в этой, совсем краткой подборке высказываний в смысле понятия «эксперимент» нет согласованности, и ни одно из этих высказываний не отвечает удовлетворительно на вопрос о том, как возможен научный эксперимент.

Не возможно всерьез воспринимать утверждение о том, что эксперимент есть предметно-чувственная деятельность, осуществляемая познанными средствами. Если, например, исследователь в процессе эксперимента имеет дело с рентгеновскими лучами, то что он предметно чувствует? Разве рентгеновский спектрограф и процесс фотографирования и проявления пленки являются теоретически познанными средствами? Вся особенность экспериментальной деятельности, в том числе и научной, заключается как раз в том, что мы, осуществляя ее, узнаем что-то новое об изучаемых явлениях, хотя и пользуемся при этом средствами, механизм действия которых остается всегда не понятным до конца.

Разве можно говорить о создании точно учитываемых условий для воспроизведения изучаемого явления? Математическая теория

эксперимента как раз и возникла из понимания того, что принципиально невозможно создать точно учитываемые условия для проведения эксперимента; результат любого эксперимента всегда связан с некоторой неопределенностью, и задача хорошей организации исследования заключается только в том, чтобы эту неопределенность минимизировать, но отнюдь не в том, чтобы ее полностью устранить.

И уж совсем не приходится говорить о физическом эксперименте, как о деятельности, направленной на наблюдение изолированного физического явления. Тогда взаимодействие явлений совсем снимается с рассмотрения, и теряется понятие о математической модели в физике и ее экспериментальной проверке.

Видимо лучше всего об эксперименте говорить, пользуясь метафорами так, как это сделал Кювье, когда сказал, что экспериментатор принуждает природу разоблачаться. Видимо лучше вовсе не пытаться давать строгое определения тому, что есть эксперимент, полагая, что это понятие не поддается компактному определению. Смысл его может стать ясным только после того, как о нем будет много сказано.

Эксперимент проводится в лабораториях, на производстве, на опытных полях и участках, в клиниках и т. д.

Эксперимент может быть **физическим, психологическим или модельным**. Он может непосредственно проводиться на объекте или на его модели. Модель обычно отличается от объекта масштабом, а иногда природой.

Если модель достаточно точно описывает объект, то эксперимент на объекте может быть заменен экспериментом на модели. Наряду с физическими моделями все большее распространение получают **абстрактные математические модели**. Можно получать новые сведения об объекте, экспериментируя на модели, если она достаточно точно описывает объект.

Эксперимент занимает центральное место в науке.

Однако возникает вопрос, насколько эффективно он используется. Джон Бернал, например, отмечал, что **научные исследования организуяются и проводятся настолько хаотично, что их коэффициент полезного действия может быть оценен величиной порядка 2%**. Для того чтобы повысить эффективность исследований, требуется нечто совершенно новое. Одним из путей является **применение математических методов, построение математической теории планирования эксперимента**.

Планирование эксперимента — это процедура выбора числа и условий проведения опытов, необходимых и достаточных для решения поставленной задачи с требуемой точностью.

При этом существенно следующее:

- стремление к минимизации общего числа опытов;
- одновременное варьирование всеми переменными, определяющими процесс, по специальным правилам — **алгоритмам**;
- использование **математического аппарата**, формализующего многие действия экспериментатора;
- выбор четкой стратегии, позволяющей принимать обоснованные решения после каждой серии экспериментов.

Задачи, для решения которых может использоваться планирование эксперимента, чрезвычайно разнообразны.

Поиск оптимальных условий, построение интерполяционных формул, выбор существенных факторов, оценка и уточнение констант теоретических моделей (например, кинетических), выбор наиболее приемлемых из некоторого множества гипотез о механизме явлений, исследование диаграмм состав—свойство — вот примеры задач, при решении которых применяется планирование эксперимента. Можно сказать, что **там, где есть эксперимент, имеет место и наука о его проведении** — **планирование эксперимента**.

Поиск оптимальных условий является одной из наиболее распространенных научно-технических задач. Они возникают в тот момент, когда установлена возможность проведения процесса и необходимо найти наилучшие (оптимальные в некотором смысле) условия его реализации.

Пусть, например, у химика возникла гипотеза о том, что при взаимодействии двух веществ должен получаться некоторый интересующий его продукт. Чтобы убедиться в правильности своей гипотезы, он начинает проводить эксперимент. Возможно, что ему повезло и он получил требуемый продукт. Однако выход продукта весьма низок, скажем, 2%. **Вот тут-то и возникает задача выбора оптимальных условий.** Требуется так подобрать концентрации реагирующих веществ, температуру, давление, время реакции и другие факторы, чтобы сделать выход возможно более близким к 100%. В данном примере находятся условия проведения процесса, оптимальные в смысле максимизации выхода требуемого продукта. Но это далеко не единственно возможная постановка задачи. Найденные условия оказались бы другими, если бы ставилась, например, цель минимизации себестоимости продукта или минимизации количества вредных примесей. Следует подчеркнуть, что

всегда необходимо четко формулировать, в каком смысле условия должны быть оптимальными. Этим определяется выбор цели исследования. ***Точная формулировка цели в значительной мере определяет успех исследования.***

Задачи, сформулированные аналогичным образом, называются ***задачами оптимизации***. Процесс их решения называется ***процессом оптимизации или просто оптимизацией***. Выбор оптимального состава многокомпонентных смесей или сплавов, повышение производительности действующих установок, повышение качества продукции, снижение затрат на ее получение — вот примеры задач оптимизации.

Эксперимент, который ставится для решения задач оптимизации, называется экстремальным. Это название связано с глубокой аналогией между оптимизацией и поиском экстремума некоторой функции.

Рассмотрим следующие две задачи.

1. Прочность бетона в значительной степени определяется маркой цемента, количеством наполнителя и количеством воды. Требуется установить связь между прочностью бетона и названными факторами.

2. Надежность некоторого полупроводникового прибора зависит от ряда технологических факторов. Требуется так подобрать значения этих факторов, чтобы надежность прибора повысилась.

Рассмотрим, какая задача из двух приведенных является экстремальной.

Чтобы облегчить выбор, укажем на признак, отличающий экстремальные задачи.

Задача является экстремальной, если цель ее состоит в поиске экстремума некоторой функции.

Чтобы установить, какая из двух задач является экстремальной, надо обратиться к их формулировкам и выяснить, где удовлетворяются требования экстремальности. В задаче 1 требуется установить связь между прочностью бетона и тремя факторами. Здесь не определено, какая прочность является оптимальной, и не требуется ее оптимизировать. В задаче 2 необходимо повысить надежность прибора. Сама постановка задачи указывает на то, что существующая надежность не удовлетворяет экспериментатора и требуется поиск таких условий, при которых ее значения повысятся. Задачи типа 1 мы будем называть ***интерполяционными***, а типа 2 — ***экстремальными***.

Определим еще ряд важных понятий, первое из которых — «***объект исследования***».

Объекты, на которых проводятся эксперименты отличаются прежде всего протекающими в них процессами.

Для описания объекта исследования удобно пользоваться представлением о кибернетической системе, которая схематически изображена на рис. 1.1. Иногда такую кибернетическую систему называют «черным ящиком» .



Рис. 1.1. Схема «черного ящика»

Стрелки справа изображают численные характеристики целей исследования. Обозначим их буквой y и назовем **параметрами оптимизации**. В литературе можно встретить другие названия: критерий оптимизации, целевая функция, выход «черного ящика» и т. д.

Для проведения эксперимента необходимо иметь возможность воздействовать на поведение «черного ящика». Все способы такого воздействия будем обозначать буквой x и называть **факторами**. Их называют также входами «черного ящика».

При решении задачи будем использовать математические модели объекта исследования. Под **математической моделью** будем понимать **уравнение, связывающее параметр оптимизации с факторами**. Это уравнение в общем виде можно записать так:

$$y=f(x_1,x_2,\dots,x_n),$$

где символ $f ()$, как обычно в математике, заменяет слова: «функция от». Такая функция называется **функцией отклика**. Нами будет рассмотрен вопрос о том, как эту функцию можно выбрать и построить. А сейчас важно понять, как получаются условия проведения опытов в том эксперименте, который мы собираемся провести.

Каждый фактор может принимать в опыте одно из нескольких значений. Такие значения будем называть **уровнями**. Может оказаться, что фактор способен принимать бесконечно много значений (непрерывный ряд). Однако на практике точность, с которой

устанавливается некоторое значение, не беспредельна. Поэтому мы будем считать, что всякий фактор имеет определенное число дискретных уровней. Это соглашение существенно облегчает построение «черного ящика» и эксперимента, а также упрощает оценку их сложности.

Все факторы должны быть управляемыми. Этого требует сама постановка планирования эксперимента, предполагающего активное вмешательство в ход эксперимента. ***Такой объект называется объектом исследования.***

Фиксированный набор уровней факторов (т. е. установление каждого фактора на некоторый уровень) определяет одно из возможных состояний «черного ящика». Одновременно это есть условия проведения одного из возможных опытов. Если перебрать все возможные наборы состояний, то мы получим полное множество различных состояний данного «ящика». Одновременно это будет число возможных различных опытов.

Чтобы узнать число различных состояний, достаточно число уровней факторов (если оно для всех факторов одинаково) возвести в степень числа факторов n : p^n , где p — число уровней. Реальные объекты, с которыми мы сталкиваемся ежедневно, обладают огромной сложностью. Так, на первый взгляд простая система с пятью факторами на пяти уровнях имеет 3125 состояний, а для десяти факторов на четырех уровнях их уже свыше миллиона!

В этих условиях мы просто вынуждены отказаться от таких экспериментов, которые включают все возможные опыты: перебор слишком велик. Тогда возникает вопрос: сколько и каких опытов надо включить в эксперимент, чтобы решить поставленную задачу? На этот вопрос можно с достаточной степенью точности дать ответ, используя методы планирования эксперимента.

Следует иметь в виду, что при планировании эксперимента не безразлично, какими свойствами обладает объект исследования. Укажем два основных требования, с которыми приходится считаться. Прежде всего существенно, воспроизводятся ли на объекте результаты эксперимента. Выберем некоторые уровни для всех факторов и в этих условиях проведем эксперимент. Затем повторим его несколько раз через неравные промежутки времени и сравним значения параметра оптимизации. *Разброс этих значений характеризует воспроизводимость результатов.* Если он не превышает некоторой заранее заданной величины (наших требований к точности эксперимента), то объект удовлетворяет требованию воспроизводимости результатов, а если превышает, то не

удовлетворяет этому требованию. Мы будем рассматривать только такие объекты, для которых требование воспроизводимости выполняется.

Планирование эксперимента предполагает активное вмешательство в процесс и возможность выбора в каждом опыте тех уровней факторов, которые представляют интерес. Поэтому такой эксперимент называется **активным**. *Объект, на котором возможен активный эксперимент, называется управляемым*. Это и есть второе требование к объекту исследования.

На практике нет абсолютно управляемых объектов. На реальный объект обычно действуют как управляемые, так и неуправляемые факторы. Неуправляемые факторы влияют на воспроизводимость эксперимента и являются причиной ее нарушения. Если требования воспроизводимости не выполняются, приходится обращаться к активно-пассивному эксперименту.

Возможно, плохая воспроизводимость объясняется действием фактора, систематически изменяющегося (дрейфующего) во времени. Тогда нужно обращаться к специальным методам планирования. Наконец, возможно, что все факторы неуправляемы. В этом случае возникает задача установления связи между параметром оптимизации и факторами по результатам наблюдений за поведением объекта, или, как говорят, по результатам **пассивного эксперимента**.

Наша цель — *изложение методов планирования экстремального эксперимента для воспроизводимых управляемых объектов*.

Планирование экстремального эксперимента — *это метод выбора количества и условий проведения опытов, минимально необходимых для отыскания оптимальных условий, т. е. для решения поставленной задачи*.

Приступая к знакомству с планированием экстремального эксперимента, надо иметь в виду, что при оптимизации распространен так называемый **детерминированный подход**, особенно широко используемый в химии. При этом предполагается построение физической модели процесса на основании тщательного изучения механизма явлений (например, кинетики, гидродинамики), что позволяет получить математическую модель объекта в виде системы дифференциальных уравнений. Несомненно, что детерминированный и **статистический** (связанный с планированием эксперимента) подходы должны разумно дополнять друг друга, а не противопоставляться, как это иногда делается.

Подведем некоторый итог изложенного выше.

Мы познакомились с основными определениями, которые используются в теории планирования экстремального эксперимента. Прежде чем приступить к эксперименту, необходимо однозначно и непротиворечиво сформулировать его цель и выбрать подходящую количественную характеристику этой цели, которую мы назвали параметром оптимизации.

Понятие «объект исследования» требует точного формального определения. Для такого определения удалось приспособить кибернетическое понятие **«черный ящик»** — **модель объекта**.

Экспериментатор, вставший на путь применения методов планирования эксперимента, должен уметь формулировать свою задачу в терминах «черного ящика».

Входы «черного ящика» называются факторами. Каждый фактор может принимать некоторое определенное число различных значений, называемых уровнями. Сочетание определенных уровней всех факторов определяет возможное состояние «черного ящика» и условия одного из возможных опытов.

Совокупность всех различных возможных состояний, определяет сложность «черного ящика» и общее число возможных опытов.

Результаты эксперимента используются для получения математической модели объекта исследования, которая представляет собой уравнение, связывающее параметр оптимизации и факторы. Такое уравнение называется **функцией отклика**.

Использование для получения модели всех возможных опытов приводит к абсурдно большим экспериментам. Задача выбора необходимых для эксперимента опытов, методов математической обработки их результатов и принятия решений — это и есть задача планирования эксперимента. Частный случай этой задачи — планирование экстремального эксперимента, т. е. эксперимента, поставленного с целью поиска оптимальных условий функционирования объекта. Планирование экстремального эксперимента — метод выбора минимального количества опытов, необходимых для отыскания оптимальных условий.

Теорией и практикой эксперимента выработаны определенные требования (условия), которым должны удовлетворять независимые и зависимые переменные. Поэтому на стадии подготовки к проведению эксперимента будут весьма полезны приведенные ниже рекомендации.

1. При выборе *выходной переменной* необходимо учитывать, что она должна иметь количественную характеристику, т. е. должна измеряться; однозначно оценивать (измерять) работоспособность объекта исследования; быть статистически эффективной, т. е. иметь

возможно меньшую дисперсию при проведении опытов (это позволяет четко различать опыты); отражать как можно более широкий спектр исследуемого явления, т. е. обладать универсальностью (практически это требование обеспечить трудно, тогда рекомендуют пользоваться так называемой обобщенной переменной); иметь достаточно четкий физический смысл. Удачный выбор выходной переменной в значительной степени определяется уровнем знания объекта.

2. При выборе *факторов* нужно выполнять следующие требования:

- фактор должен быть *регулируемым*, т. е. определенным регулирующим устройством фактор должен изменяться от значения x'_i до значения x''_i (например, расход вещества может изменяться от 30 до 40 л/ч или количество вещества A в смеси от 10 до 80%, это свойство называют *операционной определенностью*);

- точность измерения и управления фактором должна быть известна и достаточно высока (хотя бы на порядок выше точности измерения выходной переменной, очевидно, что низкая точность измерения фактора уменьшает возможности воспроизведения эксперимента);

- связь между факторами должна быть как можно меньшей (в пределе должна отсутствовать), это свойство называют однозначностью факторов, что соответствует независимости одного фактора от другого.

Ряд требований предъявляется одновременно к факторам и выходной переменной: факторы и выходная переменная должны иметь области определения, заданные технологическими или принципиальными ограничениями (пример технологического ограничения — максимальная производительность компрессора, подающего газ в реактор; пример принципиального — температура кристаллизации жидкого продукта, образующегося в результате реакции); области определения факторов должны быть таковы, чтобы при их предельных значениях значение выходной переменной оставалось в своих границах; между факторами и выходной переменной должно существовать однозначное соответствие (причинно-следственная связь).

Успех экспериментирования в значительной мере обязан теории эксперимента, которая признана дать экспериментатору ответы на следующие вопросы:

1. Как нужно организовать эксперимент, чтобы наилучшим образом решить поставленную задачу (в смысле затрат времени, средств или точности результатов)?

2. Как следует обрабатывать результаты эксперимента, чтобы получить максимальное количество информации об исследуемом объекте?

3. Какие обоснованные выводы можно сделать об исследуемом объекте по результатам эксперимента?

Как уже упоминалось, *основой теории эксперимента является статистическое представление об эксперименте* (рассматриваются случайные величины или случайные функции). Это представление отвечает действительности: как правило, итоги эксперимента связаны с некоторой неопределенностью, получающейся в результате влияния неконтролируемых факторов, случайного характера процесса на микроуровне, изменений условий эксперимента, ошибок измерения и др.

Статистическое представление об эксперименте позволило перейти к многофакторному активному эксперименту, в котором удастся надежно отделить влияние факторов от шумового фона и перейти к статистически обоснованным методам анализа результатов и принятия решений.

Теория эксперимента использует принципы и концепции кибернетики, такие как математическое моделирование, оптимизация.

Большое значение имеет концепция оптимального использования факторного пространства с точки зрения определенных характеристик точности решения. Так, удастся получать оценки коэффициентов модели, обладающие минимальной дисперсией, или удастся получать наилучшие предсказательные свойства модели.

Теория эксперимента указывает исследователю точную логическую схему и способы поиска решения задач на разных этапах исследования. Можно представить весь процесс исследования циклами, повторяющимися после решения каждой из последовательных задач исследования, причем объем знаний об объекте непрерывно увеличивается.

1.2. О научном исследовании

Понятие научного знания

Знание – идеальное воспроизведение в языковой форме обобщенных представлений о закономерных связях объективного мира.

1. *Знания* – это продукт, с одной стороны, частный (его можно присвоить), а с другой – общественный, принадлежащий всем. С другой стороны секреты производства могут быть открыты и другими людьми, чаще всего это вопрос времени. Желание прославиться также

способствует передаче знаний (открытие фосфора). Можно говорить о так называемом *неотделимом*, или неявном, знании. Речь идет о том знании, которое невозможно отделить от его носителя: индивида или научного, конструкторского и производственного коллектива. В применении к научному коллективу это означает, что для создания знания определенного уровня необходим коллектив, специалисты в котором подобраны в оптимальной квалификации.

2. *Дискретность* знания как продукта. Конкретное знание либо создано, либо нет. Не может быть знания наполовину или на одну треть.

3. Знания, подобно другим общественным (публичным) благам, будучи созданными, *доступны* всем без исключения.

4. Знания по своей природе - это *информационный* продукт, а информация после того, как ее потребили, не исчезает, как обычный материальный продукт.

90% всего количества знаний, которыми располагает человечество, получено за последние 50 лет, так же как 90% общего числа ученых и инженеров, подготовленных за всю историю цивилизации, - наши современники. Это явные признаки перехода от экономики, основанной на использовании природных ресурсов, к экономике знаний.

Функции знания:

- обобщение разрозненных представлений о закономерности природы, общества и мышления;
- хранение того, что может быть передано в качестве устойчивой основы практических действий.

*Процесс движения человеческой мысли от незнания к знанию называют **познанием**.*

***Научное знание** может быть **относительным** – являясь, в основном, верным отражением действительности, отличается некоторой неполнотой совпадения образа с объектом и **абсолютным** – достигается полное совпадение.*

Абсолютное знание не может быть опровергнуто или изменено в дальнейшем.

***Чувственное познание** (непосредственная связь человека с окружающей действительностью). Его элементами являются:*

- **ощущение** – отражение мозгом человека свойств предметов или явлений окружающего мира, которые действуют на органы чувств человека;
- **восприятие** – отражение мозгом человека свойств предметов или явлений окружающего мира в данный момент времени;
- **представление** – вторичный образ предмета или явления;

- **воображение** – соединение и преобразование различных представлений в новую картину образов .

Диалектика развития подняла человечество на новую ступень развития, когда успехи в изучении и преобразовании природы стали возможны только на пути синтеза научных знаний. Всего лишь несколько веков назад объем знаний, накопленных человечеством, был столь мал, что один человек мог не только ознакомиться со всеми основными достижениями своего времени, но и внести существенный вклад в развитие многих научных направлений. Достаточно назвать таких ученых, как Леонардо да Винчи, М.В.Ломоносов и многих других.

По мере углубления человеческих знаний и накопления информации в науке стали все более четко очерчиваться направления, изучающие те или иные конкретные особенности в окружающем нас мире, те или иные его проявления. В каждом из этих направлений стал вырабатываться свой особый язык и методы изучения. И это принесло большой успех.

Однако этот, первоначально столь обнадеживающий, путь, видимо, в недалеком будущем исчерпает себя. Дело в том, что он противоречит самой своей сути: по мере углубления знаний необходима все большая дифференциация наук, что, естественно, приводит к одностороннему освещению изучаемых явлений и объектов, поскольку из-за их многочисленности связи между отдельными областями наук ослабевают. Порождением этого процесса явился «узкий специалист», который обладает исчерпывающими знаниями только в какой-либо одной области науки. По мере развития наук и углубления знаний область науки, в которой более или менее свободно ориентируется «узкий специалист», становится все меньше. По этому поводу знаменитый математик Д. Гильберт отметил , что «узкий специалист» — *это тот, кто знает все ни о чем*. Неудивительно поэтому, что подавляющее число открытий было сделано на стыках наук. ***Природа во всех своих проявлениях едина.***

Но как усвоить хотя бы главные знания не только «своей», но и смежных наук? Ведь объем имеющейся на сегодня информации по каждой из конкретных наук и даже весьма узких ее областей огромен. Однако эта задача не столь безнадежна, как это может показаться вначале. Еще Г. Галилей отмечал: «Философия написана в грандиозной книге — Вселенной, которая открыта нашему пристальному взгляду. Но понять эту книгу может лишь тот, кто научился понимать ее язык и знаки, которыми она изложена. Написана

же она на языке математики...». Один из путей решения ее — **широкая компьютеризация процессов обучения и обеспечения научно-технической информацией.**

По мере развития науки результаты научных исследований оказываются необычайно важными, а иногда и жизненно важными для развития целых отраслей промышленности, поэтому все более отчетливо выступает еще одна тенденция — стремление к возможно более быстрому внедрению в промышленность результатов научных исследований. И успехи на этом пути впечатляющи: на разработку средств телефонной связи потребовалось примерно 56 лет, радио — 35 лет, радиолокации — 12 лет, транзисторов — 5 лет, а интегральных схем — всего 3 года.

Подготовка современного специалиста — это длинный и сложный процесс, с течением времени отнимающий все большее количество лет жизни человека, процесс, требующий огромных затрат и духовных и материальных, и, тем не менее, он оказывается оправданным, если заканчивается пополнением науки и производства высококвалифицированными кадрами.

О значении квалифицированного специалиста в хозяйственной деятельности общества свидетельствует такой случай. Одна крупная фирма, занимающаяся производством громадных дорогостоящих электрогенераторов, обратилась к известному электротехнику-практику Штейнметцу с просьбой помочь разобраться в том, почему последний из созданных этой фирмой электрогенераторов выдает лишь половину мощности по сравнению с расчетной. Ознакомившись с представленной информацией, Штейнметц внимательно прослушал работающий генератор, после чего попросил лестницу. Приставив ее к корпусу генератора, он поднялся по ней и, просверлив дрелью в корпусе отверстие, пересоединил несколько проводов, после чего генератор стал выдавать положенную ему мощность.

Спустя некоторое время Штейнметц прислал фирме счет за свою работу на 10 000 долларов. Фирма в суде опротестовала его: десять тысяч за просверленную дырку — это ни в какие рамки не укладывается!

На вопрос о том, за что он взял эти десять тысяч долларов, Штейнметц ответил так:

— За то, что просверлил дырку и пересоединил провода - 1 доллар; за то, что знал, как исправить генератор, — 9 999 долларов.

Под **научным исследованием** понимается процесс изучения объекта или явления с целью выявления закономерностей его

возникновения и развития, а также возможностей изменения и преобразования его в интересах общества.

Разработка наиболее эффективных приемов и средств для проведения научных исследований всегда занимала умы специалистов, работающих в самых различных областях естествознания. Правила, сформулированные для решения математических задач, не могут быть без изменений перенесены на задачи другого типа. Более того, различные сферы деятельности всегда будут в определенной мере специфичными.

Понятие научного исследования не следует путать с понятием испытания. Хотя в одной и той же работе эти понятия могут идти рядом, их надо разграничивать. Под **испытанием** понимают проверку каких-либо установленных заранее (ГОСТ, ТУ, нормами и требованиями министерств) свойств или качеств объекта. Например, при выпуске двигателя проверяют, развивает ли он гарантированную мощность, каков при этом расход топлива и т. п. Но если ставится задача, каким образом наиболее целесообразно повысить мощность этого двигателя и каковы при этом будут последствия при эксплуатации двигателя с повышенной мощностью, — это уже предмет исследований.

Исследование может быть теоретическим или экспериментальным. **Научная абстракция** — *общий метод теоретического исследования.* В приложении к технике теоретическое исследование заключается в анализе и синтезе закономерностей и открытии новых закономерностей при помощи математического аппарата. Используя известные математические выражения законов природы, устанавливают и объясняют новые объективные закономерности путем логического развития, принятия новых предпосылок и допущений.

Теоретическое исследование обладает свойствами всеобщности (для данной группы явлений или их развития в определенных пределах). Оно не требует больших затрат труда и средств. При достаточной надежности исследования его результаты действительны в течение длительного времени. Правильность теоретического исследования и его надежность (с точки зрения соответствия закономерностям объективного мира) зависят от правильности предпосылок, допущений и логического развития.

Изучаемые явления настолько многообразны и сложны, что часто трудно, а порою и невозможно получить точное теоретическое решение, поэтому результаты расчетов становятся ненадежными.

Теория, или научно обобщенная система знаний, должна соответствовать практике, иначе принятые исследователем положения

и построения в лучшем случае будут гипотетическими, предположительными, а иногда и просто досужим вымыслом. Очень часто только эксперимент (научно поставленный опыт) является единственным надежным способом решения поставленной задачи.

Роль экспериментальных исследований в развитии науки огромна. Они дают исходные данные для дальнейшего развития существующей теории или построения новой и помогают контролировать жизненность теоретических выводов. В этом смысле опыт—процесс изменения природы человеком, основа познания и критерий истины: человек может познавать природу, лишь воздействуя на нее.

Опытное решение — основа экспериментального исследования; оно (решение) всегда конкретно и точно (в пределах, обусловленных применяемой методикой), но иногда громоздко и трудоемко.

Практика выше теоретического познания, так как она имеет не только достоинство всеобщности, но и непосредственной действительности. Но экспериментальное исследование полностью этой всеобщности не отмечает и не может установить всю бесконечную сложность взаимосвязанных связей, характеризующих сущность явления. Мы по необходимости выбираем для изучения и открытия природных закономерностей лишь какую-то часть сущности. Таким образом, экспериментальное и теоретическое исследования дополняют друг друга, расширяют и углубляют на каждом своем этапе наши знания о развитии явления.

В большинство случаев нельзя обособлять теоретические и экспериментальные исследования, так как в основе теоретического исследования лежит опыт, а обобщение опытных данных развивает теорию.

Исследование, состоящее из экспериментальной и теоретической части, называется комплексным.

Под экспериментальным исследованием обычно подразумевается комплексное.

Термин «**комплексное исследование**» может иметь и несколько иное понятие. Исследование какого-либо явления может проводиться специалистами различного направления; в сельском хозяйстве, особо ценны исследования, проводимые с одной общей целью биологами, инженерами и экономистами, причем каждый из специалистов исследует экспериментально и теоретически свое направление, так что в общем получается многостороннее исследование поставленной задачи. Такие комплексные исследования обычно весьма эффективны, и их следует развивать.

По организации исследования разделяют на лабораторные, заводские (полевые, клинические и др.) и лабораторно-заводские.

В какой бы конкретной области ни проводились исследования, им присущи некоторые общие черты, а именно компоненты исследования, которые можно сформулировать так:

1. Постановка задачи, проблемы.
2. Предварительный анализ имеющейся информации, условий и методов решений задач данного класса.
3. Выдвижение гипотез, новых научных и технических решений.
4. Планирование и организация натурального или теоретического эксперимента.
5. Проведение эксперимента.
6. Анализ и обобщение полученных результатов.
7. Проверка исходных гипотез на основе полученных фактов.
8. Окончательная формулировка новых фактов и законов. Получение объяснений и предсказаний.

Это, разумеется, лишь общая схема исследований, которая может в той или иной мере видоизменяться в зависимости от имеющихся исходных данных и от конкретных целей исследования.

Заключительным этапом любого прикладного исследования должно быть внедрение в производство полученных результатов.

Научные исследования могут проводиться на двух взаимосвязанных уровнях:

- 1) **эмпирическом** (нахождение новых фактов и формулирование на их основе анализа, синтеза и обобщения эмпирических закономерностей);
- 2) **теоретическом** (формулирование общих для данной предметной области закономерностей, на основе которых глубже интерпретируются как новые факты, так и полученные на эмпирическом уровне закономерности, создается возможность прогнозирования дальнейшего процесса развития исследуемой предметной области).

Каждое научное исследование в явном или неявном виде включает в себя:

1. **Всеобщие философские методы** (методы диалектического материализма — принцип всеобщей взаимосвязи, развития, движения и изменения; принцип единства и борьбы противоположностей; методы, вытекающие из знания законов диалектики — единства и борьбы противоположностей, перехода количественных изменений в качественные и обратно, отрицания отрицания; диалектических категорий — «сущность и явление», «форма и содержание», «возможность и действительность» и др.).

2. **Общие научные методы** (анализ и синтез, индукция и дедукция, аналогия, аксиоматизация, формализация, математизация и т. д.)

3. **Частные методы**, специфически характерные для данной конкретной области науки. При этом конкретный способ исследования, подборка используемых методов меняются вместе с изменением предмета исследования.

В начале любого исследования надо **определить цель и выбрать предмет исследования**. Эти действия не могут быть случайными: научное исследование — одна из основ развития общества, оно должно содействовать осуществлению планов и задач стоящими перед обществом. Установление законов развития явления неотделимо от их общественной полезности.

Определив цель и выбрав предмет исследования, исполнитель должен полностью изучить материалы исследований, проведенных ранее в этом же направлении, и критически проанализировать их. Этот этап, иногда довольно обширный, исключает поспешные выводы, неправильные заключения и повторение уже сделанного.

Следующий этап — **составление рабочей гипотезы исследования** (предположения о вероятном развитии явления), которая в наибольшей степени устраняет из исследования случайные поиски решения, определяет направление исследования и его границы.

Общим методом (путем, способом, приемами) любого исследования, как процесса познания окружающего нас мира, является материалистическая диалектика. Не зная диалектического метода, трудно проникнуть в сущность явления и использовать изучаемые закономерности.

В научных исследованиях используются различные методы, присущие той или иной отрасли науки, однако один из общенаучных методов используется всегда. Речь идет о **математизации во всех компонентах научного исследования**.

Математика в научных исследованиях — это вовсе не искусственный язык, придуманный людьми, и не просто сокращенный способ записей стенографического типа; это **естественный язык науки, и никакого другого сегодня мы не знаем**.

Для того чтобы оценить роль математики, нужно ответить на вопросы: Что такое математический метод для исследователя? В чем наиболее ярко проявляется его сущность? Прежде всего, в так называемом дедуктивном подходе, так что каждая дедуктивная теория может считаться математикой.

После того как **сформулированы аксиомы и определения**, целью **математического рассуждения, основанного на законах логики**,

является выявление тех следствий, которые вытекают из первоначальных посылок. **Первоначальные же посылки — аксиомы (постулаты) и частично определения** — обусловлены теми опытными данными, которые имеются в распоряжении исследователя, с той его проницательностью, которая позволяет заглянуть вглубь природных явлений. Своеобразие математического подхода начинает проявляться с вводимых определений, которые, по существу, сводятся к тому, что вместо некоторой комбинации известных символов используется один новый. Это позволяет значительно сократить формулировки как посылок, так и алгоритма получения следствий из них, которые в противном случае могли бы быть попросту необозримыми.

Таким образом, наиболее ответственным этапом применения математических методов становится формулировка исходных посылок.

Исходные посылки — это некоторые утверждения, которые, как правило, не могут быть абсолютными. Поэтому вытекающие из них следствия справедливы лишь в определенных границах. Эту особенность любых теоретических результатов хорошо понимают экспериментаторы по самому роду своей работы.

Возможности же современной математики огромны. Проиллюстрируем это на примере.

Положим, изучается новое физическое явление. Добиться успеха возможно, только научившись описывать его и предсказывать его поведение в различных ситуациях. **Это возможно лишь тогда, когда будет построена математическая модель для описания явления**, которая в общем случае связывает множество некоторых характеристических параметров $U = (U_1, U_2, \dots, U_m)$ и множеством других величин $f = (f_1, f_2, \dots, f_m)$ с помощью, как правило, непростой зависимости $A(U) = f(\bullet)$, записанной в векторной форме, в которой $A = (A_1, A_2, \dots, A_m)$ — вектор операторов в m -мерном пространстве.

Спрашивается, а насколько разумна построенная модель? Допускает ли она решение? Для ответа на такой вопрос существует одна из очень полезных в математике теорем — **теорема о неподвижной точке**. В одном из вариантов она звучит так: **непрерывное отображение замкнутого ограниченного выпуклого множества M евклидова пространства R_m в себе имеет по меньшей мере одну неподвижную точку**.

С ее помощью удастся решать огромное число задач из самых различных областей естествознания. Если кому-нибудь покажется, что она слишком абстрактна, то этот человек глубоко заблуждается. Вот пример, опровергающий это заблуждение. Возьмем стеклянный шар,

заполненный мелким песком, и встряхнем его. Интуитивно кажется, что все песчинки изменят свое положение. Но это неверно! Хотя бы одна из них останется неподвижной — это следствие сформулированной теоремы.

В данном случае при известном $f(\bullet)$ роль отображения играет совокупность значений $A(U)$. Если каждая зависимость $A_n(U)$ непрерывна относительно аргументов U_i , а это нетрудно проверить, то первое условие теоремы — непрерывность отображения — будет выполнено.

Значения каждой из величин U_i и f_n заключены, очевидно, в некотором интервале значений, совокупностью которых является множество M . Чаще всего каждый из интервалов вполне можно считать ограниченным; тем самым будет выполнено условие замкнутости.

Если значения величин U_i и U_k (или f_i и f_p) находятся на множестве M ($U_i, U_k \in M$), то и $[(1 - \lambda) U_i + \lambda U_k] \in M$, где $0 \leq \lambda \leq 1$ (аналогично $[(1 - \lambda) f_i + \lambda f_p] \in M$). Иначе говоря, возможные значения U_i «сплошь» заполняют множество M ; наряду с элементами U_i и U_k множеству M принадлежит и весь отрезок $(1 - \lambda) U_i + \lambda U_k$, соединяющий U_i и U_k . Тем самым будет выполнено условие выпуклости множества M . При соблюдении отмеченных условий построенная система уравнений, очевидно, имеет решение. На этом этапе возникает другая проблема: как найти это решение? Чтобы убедиться в правильности решения, можно использовать хорошо разработанные методы вычислительной математики, например прямые методы вариационного исчисления.

Конечно, можно указать много примеров, свидетельствующих о том, что математическая идеализация лишь в определенных пределах согласуется с содержанием физической действительности. Но если это обстоятельство не упускать из виду, то математический метод выступает со всей своей силой как эффективнейшее и незаменимое средство при проведении научных исследований.

Современные средства исследования приспособлены к точной регистрации отдельных сторон развивающегося явления, к установлению связи между этими сторонами, но они не могут исключить ошибку в обобщении, в выражении находимого закона или в построении математической модели явления. Поэтому тщательность исследования, обязательная и неоднократная проверка суждения, правильный выбор математического аппарата исключительно важны.

Рассматривая основные вопросы мышления при проведении исследований, не следует пренебрегать интуицией исследователя.

Под **интуицией** следует понимать способность человеческого ума к чрезвычайно ускоренному процессу логического мышления, так что интуитивно найденное решение не только постороннему, но и самому исследователю может показаться мгновенным, внезапным.

Способность к интуитивному мышлению присуща в той или иной степени каждому человеку и может развиваться упражнениями. Но для проявления интуиции необходимы: достаточно полная информация об обсуждаемом явлении; напряженная, непрерывная работа мышления; при всем этом обдумывание решения с разных точек зрения, в разных направлениях. Творческое воображение, поиск разрешения противоречий, анализ и синтез последних в их диалектическом единстве — таковы общие признаки проявления интуитивного мышления. Конечно, интуитивное решение требует не только тщательного логического развития, но и экспериментальной проверки, так как примеры неправильных интуитивных решений, обусловленных чаще всего недостаточной информацией, у молодых исследователей пока еще имеют место.

Приемы, способы исследования зависят от его цели, применяемых средств и особенностей других условий, в которых оно должно проводиться. В каждой науке эти приемы имеют общие черты, что позволяет применять общую методику экспериментального исследования. Любое исследование должно быть планомерным по своей сущности и проводиться по определенной методике.

Итак, прежде чем начать экспериментальное исследование , необходимо:

- определить цель и выбрать предмет исследования;
- освоить накопленные знания и критически проанализировать их;
- составить методику и план исследования.

1.3. Методология теоретических и экспериментальных исследований

1.3.1. Законы и формы мышления (мышление, понятие, абстракция)

Мышление – это процесс опосредствованного отражения действительности. Благодаря абстрактному мышлению мы получаем новые знания не непосредственно, а через посредство других знаний.

Знание, полученное из уже имеющихся знаний без непосредственного обращения к опыту (практике), называют **выводным**, а сам процесс его получения – **выведением**.

Выведение знаний осуществляется в процессе **рассуждений**, которые должны подчиняться законам мышления.

Мышление как средство познания объективного мира изучает наука логика.

Определенность и последовательность мышления невозможны без точного употребления понятий.

Понятие – это результат отражения в сознании человека общих свойств группы предметов или явлений, которые существенны и необходимы для выделения рассматриваемой группы.

Понятие - общее, единичное, собирательное (относящееся к группам предметов - лес, транспорт), конкретное, абстрактное (к отдельно взятым признакам предметов - белый), относительное – парное (правый-левый, начальник-подчиненный), абсолютное – не имеет парных отношений (дом, дерево).

Свойства и отношения, характеризующие тот или иной объект, называются **признаками**.

Признаки, которые выражают внутреннюю природу объекта, его сущность, называются **существенными**. Они всегда принадлежат данному объекту.

Признаки, которые могут принадлежать, но могут и не принадлежать объекту, и которые не выражают его сущности, называются **несущественными**.

Кроме того, признаки разделяются на **отличительные** и **неотличительные**.

Отличительные признаки присущи рассматриваемому объекту (или определенному классу объектов) и позволяют выделить его (их) из всего многообразия объектов.

Неотличительные признаки могут принадлежать не только рассматриваемому объекту, но и другим. Например, кратные связи у непредельных углеводородов, но у алкенов одна двойная связь, а у алкинов – тройная.

Абстракция – это метод научного исследования, основанный на отвлечении от несущественных сторон и признаков рассматриваемого объекта. Абстракция позволяет упростить объект или процесс, заменить его моделью, т.е. другим эквивалентным в определенном смысле объектом или процессом (исходя из условий задачи) и исследовать эту модель.

Различают **три типа абстракции**.

Изолирующая абстракция производится для вычленения и четкой фиксации исследуемого объекта по существенным признакам.

Обобщающая абстракция применяется для получения общей картины процесса или явления. Например, в результате обобщения свойств электрических, пневматических, гидравлических машин, жидкостных реактивных двигателей, двигателей внутреннего сгорания возникает такая обобщающая абстракция как преобразователь энергии. Работу парового двигателя, двигателя внутреннего сгорания, ракетного двигателя, холодильника можно рассматривать с единых позиций термодинамики как работу тепловой машины.

Идеализирующая абстракция заключается в замещении реального объекта идеализированной схемой для упрощения процесса его изучения. При идеализации объектов необходимо четко сформулировать принятые допущения.

Например, при расчете конструкции на прочность реальные шарнирные опоры заменяют идеальными, считая, что трение в опорах отсутствует. Следствием идеализации модели может стать превышение напряжений, действующих в реальной конструкции, над расчетными значениями. Поэтому в расчеты вводят коэффициенты безопасности.

Идеализирующая абстракция используется при мысленном конструировании понятий о несуществующих и, может быть, неосуществимых объектах, но имеющих прообразы в реальном мире. Например, точка (*в реальном мире нет объекта, который не имеет измерений*), прямая, инерция, абсолютно черное тело и др. Созданные идеальные объекты находят определенное истолкование в терминах реальных объектов и их применение подтверждено практикой научно-технического развития.

1.3.2. Законы и формы мышления (сравнение, индукция и дедукция, анализ и синтез)

Сравнение – это операция мышления, направленная на установление сходства или различия изучаемых объектов по каким-либо признакам. В основе операции лежит классификация сравниваемых понятий.

Операция сравнения может выполняться только для однородных объектов, входящих в определенный класс. Формирование такого класса объекта, а также определение состава существенных и отличительных признаков сравнения в ряде случаев представляет собой достаточно сложную интеллектуальную задачу.

Индукция (лат. *induction* – наведение) – операция мышления, основанная на обобщении эмпирической информации об устойчивой повторяемости признаков ряда явлений.

Индуктивные умозаключения позволяют от отдельных фактов перейти к общему знанию.

Индуктивные умозаключения в большей степени способствуют получению новых знаний. История науки показывает, что многие научные открытия в физике, химии, биологии сделаны на основе индуктивного обобщения эмпирических данных.

В зависимости от полноты и законченности эмпирического исследования различают *полную* и *неполную* индукцию.

При *полной индукции* на основе повторяемости признаков у каждого явления (объекта), относящегося к определенному классу, заключают о принадлежности этого признака всему классу. Это возможно в тех случаях, когда исследователь имеет дело с замкнутыми классами, число элементов (объектов) в которых является конечными и легко обозримыми.

При *неполной индукции* на основе повторяемости признака у некоторых явлений, относящихся к определенному классу, заключают о наличии этого признака у всего класса явлений. При этом подразумевается, что сам класс сформирован по каким-либо другим признакам, а не тем, что анализируются.

Логический переход в неполной индукции от некоторых элементов ко всем элементам класса не является произвольным. Он оправдан устойчивыми эмпирическими основаниями. Однако, обобщение в этом случае носит вероятностный характер, и вывод может содержать ошибки. Например, большинство сталей и сплавов имеют положительный коэффициент термического расширения, причем значительно больший, чем у неметаллов. Но обобщающего вывода сделать нельзя, например, сплав инвар марки И-36, содержащий 36% Ni, при температуре от -50 до 100°C имеет коэффициент линейного расширения, близкий к нулю.

Дедуция (лат. *deduction* – выведение) – операция мышления, заключающаяся в том, что на основании общего знания выводятся частные положения. Дедуктивные умозаключения обладают высокой степенью доказательности и убедительности.

Дедуктивные рассуждений (от известных общих закономерностей) могут приводить к эффективным частным решениям. Например, известно, что усталостное разрушение конструкции от внешних нагрузок происходит в результате зарождения трещин в поверхностном слое. Трещины появляются в результате действия растягивающих напряжений. Отсюда вывод – если при изготовлении детали в поверхностном слое создать внутренние сжимающие напряжения, то можно повысить усталостную прочность конструкции.

Анализ (греч. analysis – разложение, расчленение) – процедура разложения объекта (предмета, явления, процесса) на составные части.

Особую специфику представляет анализ технического объекта (ТО). При анализе ТО можно выделить два подхода:

1. Мысленное или реальное расчленение объекта на составные элементы. При этом выявляется структура объекта, т.е. состав элементов и отношения между ними, исследуются причинно-следственные связи между элементами.

Например, космический аппарат (КА) можно рассматривать как совокупность систем – системы двигательной установки, системы ориентации КА, управления научной аппаратурой, системы терморегулирования и др. Каждая система анализируется как автономный комплекс объектов определенного функционального назначения.

Используя методы абстракции, можно описать элементы системы при помощи идеализированных моделей, определить оптимальные параметры каждой системы;

2. Расчленение свойств и отношений объекта на составляющие свойства и отношения. При этом одни из них подвергаются дальнейшему анализу, а от других отвлекаются. Затем подвергаются анализу те свойства, от которых отвлекались. В результате понятия о свойствах и отношениях исследуемого объекта сводятся к более общим и простым понятиям. Изолирующая абстракция является частным случаем такого анализа.

Примером может служить анализ трубопроводной системы, с одной стороны, как объекта, обладающего определенным гидравлическим сопротивлением, а с другой – как объекта, который не должен разрушаться при действии на него различных нагрузок.

Синтез (греч. synthesis – соединение, сочетание, составление) – метод научного исследования какого-либо объекта, явления, состоящий в познании его как единого целого, в единстве и взаимной связи его частей.

Синтез, с одной стороны, является методом познания, с другой – это метод практической деятельности. Процессы проектирования, конструирования определяются как операции синтеза. При этом новый полученный объект имеет существенно другое качество, чем элементы его составляющие. Это не сумма элементов, это более сложное взаимодействие.

Синтез является приемом, противоположным анализу. Вместе с тем оба приема предполагают и дополняют друг друга.

Без анализа нет синтеза, без синтеза – анализа.

Например, при разработке космического аппарата как комплекса систем, анализ каждой системы и оптимизация ее параметров сопровождается исследованием совместной работы всех систем с учетом их взаимодействия.

1.3.3. Законы и формы мышления (обобщение, аналогия, гипотеза)

Обобщение – это операция мышления, заключающаяся в переходе от частного к общему, причем на более высокую ступень абстракции. Индукция является частным случаем обобщения. Обобщение позволяет формулировать общие принципы и законы на основе исследования частных явлений.

По семантико-гносеологическому (смысловому, познавательному) содержанию обобщения подразделяются на два вида:

1. Обобщения, порождающие новые понятия, формулирующие законы, принципы, теории, которые не определяются исходным смысловым содержанием изучаемых явлений. Это обобщения, позволяющие выявить общую сущность по-разному воспринимаемых явлений.

Например, обобщением понятий «радиоволны», «тепловые лучи», «свет», «гамма-излучение» является понятие «электромагнитные колебания, отличающиеся различной длиной волны»;

2. Обобщения, позволяющие применить известные явления, принципы, закономерности, действующие в одной области знаний, в другой области.

Например, использование данных о форме тела рыб, дельфинов, китов в судостроении с целью создания судов с минимальным гидравлическим сопротивлением.

Аналогия является одним из способов логического перехода от известного к новому знанию, выдвижения предположений.

Аналогия – это умозаключение о принадлежности объекту определенного признака (т.е. свойства и/или отношения) на основе сходства в существенных признаках с другим объектом.

В зависимости от характера связи между признаками различают строгую аналогию, дающую достоверное заключение, и нестрогую аналогию, дающую вероятностное заключение.

На принципах строгой аналогии построены доказательства в области математики.

При нестрогой аналогии зависимость между сходными и переносимыми признаками определена как необходимая с определенной степенью вероятности. Поэтому, обнаружив у другого объекта признаки сходства, можно лишь в вероятностной форме утверждать о наличии у него переносимого признака.

Условиями, повышающими степень вероятности выводов в нестрогой аналогии, является:

1. Сходство употребляемых объектов в значительном числе существенных признаков.
2. Отсутствие существенных различий между употребляемыми объектами.
3. Высокая вероятность знания о зависимости между сходными и переносимыми признаками.

При решении научных и инженерно-практических задач обращаются не только к известным законам и научным обобщениям, но и к ранее приобретенным знаниям о сходных единичных явлениях.

Аналогия используется не только для поиска конкретного инженерного решения, но и при выборе метода решения поставленной задачи.

Гипотеза – это форма развития человеческих знаний, представляющая собой обоснованное предположение, объясняющее свойства и причины исследуемых явлений.

Достоверному познанию явлений объективного мира всегда предшествует длительная работа по осмыслению эмпирического материала. Эта работа всегда сопровождается построением различного рода догадок, предположений относительно действительных причин наблюдаемых явлений, преодолением различного рода противоречий. Результатом этой работы является гипотеза, без которой невозможен переход от незнания к знанию, т.е. *гипотеза – это форма развития знания (получения знания).*

Новое знание первоначально всегда возникает в форме гипотезе, построение которой начинается с индуктивного обобщения эмпирического материала. При этом могут использоваться дедуктивные выводы, методы аналогий.

На основе обобщения выдвигаются предположения.

В отличие от предположений гипотеза должна быть состоятельной, т.е. удовлетворять определенным логико-методологическим требованиям. Она должна быть непротиворечивой (т.е. не противоречить всему исходному эмпирическому материалу), принципиально проверяемой, эмпирически и теоретически обоснованной (т.е. иметь

достаточную предсказательную и объяснительную силу для отыскания новых фактов и их объяснения).

В зависимости от объекта исследования различают гипотезы общие и частные:

Общая гипотеза – это научно обоснованное предположение о закономерностях естественных и общественных явлений. Они выдвигаются для объяснения всего класса описываемых явлений. **Будучи доказанными, они становятся научными теориями и являются ценным вкладом в развитие научных знаний.**

Частная гипотеза – это научно обоснованное предположение о причинах, происхождении и закономерностях функционирования группы объектов, выделенных из класса рассматриваемых.

Гипотеза считается доказанной и переходит в разряд достоверных истин, во-первых, если предположение, составляющее содержание гипотезы, выводится, как следствие, из общего характера (т.е. доказывается методом дедукции), во-вторых, если путем наблюдения существующих объектов и процессов или проведением специального эксперимента могут быть практически обнаружены новые факты, подтверждающие содержание гипотезы. В последнем случае гипотеза дает направление поиска этих фактов и позволяет правильно организовать проведение эксперимента.

1.3.4. Методология исследований

Методология – учение о методах познания и преобразования действительности.

Методологией руководствуется движение мысли от незнания к знанию.

У методологии две функции:

Общая – обоснование правил применения мировоззрения к процессу познания и преобразования мира,

Частная – определение подхода к явлениям действительности.

Метод – способ достижения цели.

Методы подразделяют на несколько уровней:

- **эмпирический уровень**, на нем применяют наблюдение, сравнение, счет, измерение и др., при этом происходит накопление фактов и их описание;

- **экспериментальный** (теория, гипотеза) – эксперимент, анализ-синтез, индукция-дедукция, моделирование, логический метод. На этом уровне осуществляется также описание-накопление фактов и их проверка. Факты имеют ценность, только когда они систематизированы, проверены, обработаны;

- **теоретический** – абстрагирование, идеализация, формализация, анализ-синтез, индукция-дедукция, аксиоматика, обобщение. На этом уровне проводится логическое исследование собранных фактов, выработка понятий, суждений, умозаключений. Соотносятся ранние научные представления с возникающими новыми, создаются теоретические обобщения. Новое теоретическое содержание знания надстраивается над эмпирическими знаниями;

- **метатеоретический** – это диалектический метод и метод системного анализа. Этими методами используются сами теории, разрабатываются пути из построения, устанавливающие границы из применения. Т.е. **на этом уровне происходит познание условий формализации научных теорий и выработка формализованных языков, именуемых метаязыками.**

1.3.5. Задачи теоретических исследований

Цель – выявление существующих связей между исследуемым объектом и окружающей средой, объяснение и обобщение результатов эмпирических исследований, выявление общих закономерностей и их формализация.

В процессе теоретического исследования приходится непрерывно ставить и решать разнообразные по типам и сложности задачи в форме противоречий теоретических моделей, требующих разрешения.

В логико-психологическом аспекте **задача** – это несогласованные или противоречивые информационные процессы (системы), соотношение между которыми вызывает потребность в их преобразовании. В процессе решения задачи противоречия между указанными информационными процессами или системами устраняются.

Структурно любая задача включает условия и требования.

Условия – это определенная информационная система, из которой следует исходить при решении задачи.

Требования – это цель, к которой нужно стремиться в результате решения.

Условия и требования могут быть исходными, привлеченными и искомыми.

Исходные – даются в первоначальной формулировке задачи (исходные данные). Если их недостаточно для решения, то привлекают новые – **привлеченные**.

Искомые – это те, которые требуется отыскать в процессе решения задачи.

Типы теоретических задач:

- обобщение результатов исследований, нахождение общих закономерностей путем обработки и интерпретации опытных данных;
- расширение результатов исследований на ряд подобных объектов без повторения всего объема исследований;
- изучение объекта, недоступного для непосредственного исследования;
- повышение надежности экспериментального исследования объекта (обоснования параметров и условий наблюдения, точности измерений).

Для решения задач используют анализ и синтез, а также расчленение и объединение элементов исследуемой системы (объекта, явления).

Метод **декомпозиции** (расчленения) – выделяют существенные и несущественные параметры, основные элементы и связи между ними, строится модель объекта исследования.

Метод **объединения** (общая теория систем,). Некоторые постулаты общей теории систем:

1. Функционирование систем любой природы может быть описано на основе рассмотрения формальных структурно-функциональных связей между отдельными элементами систем. Влияние материала элементов системы проявляется в формальных характеристиках системы (структуре, организации, динамике);

2. Организация системы может быть определена на основе наблюдений, проведенных извне посредством фиксирования только тех элементов системы, которые непосредственно взаимодействуют с ее окружением.

1.4. Цель и предмет исследования

1.4.1. Цель исследования

Выбор цели научных исследований — это один из важнейших творческих этапов в решении различных тем и проблем. Так, при разработке новых радиоэлектронных систем выбрать цель — значит решить вопрос: систему какого класса и для какого назначения следует разрабатывать. Во многих случаях цель задается разработчиком, главным конструктором, иногда цель диктуется результатами процесса творчества, например «мозговым штурмом».

В любой ситуации принятия того или иного решения исследователь (разработчик) должен прежде всего ответить на вопрос: к чему он стремится? В чем состоит цель исследования? Вряд ли можно дать четкий ответ на этот вопрос при слишком «глобальном» подходе, например, к цели существования всего человечества, Солнечной

системы и т. п. Однако в рамках четко очерченной ситуации ответ может быть точно определенным. В математических моделях принятия решений существует способ формализации цели, состоящий во введении **целевой функции**. Такая функция для каждого возможного исхода численно оценивает его полезность для лица, принимающего решение, хотя не всегда, удобна, так как не все сводится к числу или системе чисел. Так, нельзя оценить труд ученого только количеством опубликованных им работ. Можно опубликовать одну работу, но решить в ней чрезвычайно крупную проблему в той или иной области науки, перевернуть существующие до этого представления, сделать открытие. Понятно, что здесь суть не в количестве научных работ.

В реальных задачах мы обычно имеем дело со сложными системами: техническими, экономическими, биологическими, экологическими, социальными и т. д. Как же определять целевую функцию в тех случаях, когда формальное описание может быть нереализуемым? **Под формальным описанием понимается математическая (знаковая) модель объективной действительности.** Математическая модель задачи принятия решения — формальное описание составляющих ее компонентов: **цели, средств, результатов, а также связи между последними.** Можно дать формальное описание цели, взяв за основу связанное с ней отношение предпочтения. Обычно для этого выделяют множество всех тех пар результатов, для которых один результат соответствует цели ближе, чем другой (при двоичном выборе: «который из двух»). Здесь легко составить отношение предпочтения, так как в такой ситуации не требуется проведения количественной оценки.

Для более эффективной работы над научной проблемой, разрабатывают «дерево» целей. Рассмотрим разработку дерева целей на примере НИР и ОКР в области радиоэлектроники (рис. 1.2).



Рис. 1.2. «Дерево целей» НИР и ОКР в области радиоэлектроники

При разработке дерева целей необходимо исходить из общих требований: соблюдения иерархии целей согласно системных методов; относительной стабильности цели для рассматриваемого периода времени; полноты совокупности целей; отражения в дереве целей как теоретического, так и прикладного аспектов исследования.

Перечислим основные уровни и подцели повышения эффективности радиоэлектронного производства :

Уровень О

Повышение эффективности радиоэлектронной аппаратуры (генеральная цель развития радиоэлектроники).

Уровень А

A₁. Повышение помехоустойчивости радиоэлектронных систем.

A₂. Повышение пропускной способности систем передачи информации и каналов связи.

А₃. Повышение надежности радиоэлектронной аппаратуры (РЭА) — конструктивной, технологической.

А₄. Обеспечение требований электромагнитной и экологической совместимости РЭА.

А₅. Освоение новых диапазонов электромагнитных и акустических волн, обеспечение дальности действия.

А₆. Развитие элементной базы, микроминиатюризация.

Уровень В

Подцель А₁

А₁В₁. Совершенствование методов оптимального радиоприема, применение цифровых систем.

А₁В₂. Совершенствование методов помехоустойчивого кодирования и декодирования.

А₁В₃. Совершенствование методов обработки сигналов и сообщений, одновременная оптимизация модемов и кодеков.

А₁В₄. Совершенствование возбuditелей и радиопередающих устройств, улучшение фильтрации гармоник и побочных излучений.

Подцель А₂

А₂В₁. Совершенствование методов уплотнения при многоканальной передаче сообщений.

А₂В₂. Применение методов сокращения избыточности систем.

А₂В₃. Совершенствование методов сжатия спектра частот.

А₂В₄. Повышение скорости передачи сообщений, адаптивная коррекция каналов связи.

Подцель А₃

А₃В₁. Совершенствование методов последовательного резервирования блоков РЭА.

А₃В₂. Совершенствование методов параллельного резервирования блоков РЭА.

А₃В₃. Совершенствование конструкции и технологии изготовления РЭА.

Подцель А₄

А₄В₁. Совершенствование антенно-фидерных устройств.

А₄В₂. Совершенствование способов экранирования отдельных блоков и узлов РЭА.

А₄В₃. Разработка новых методов (фильтров) помехозащиты.

А₄В₄. Оптимизация формы передаваемого сигнала, совершенствование методов оптимального радиоприема.

Подцель А₅

А₅В₁ Совершенствование конструкции и технологии средств канализации электромагнитной энергии (волноводов, световодов).

А₅В₂. Совершенствование методов волоконной оптики.

А₅В₃. Совершенствование и развитие акустоэлектронных систем.

Подцель А₆

А₆В₂. Совершенствование технологии и конструкции БИС, увеличение степени интеграции микрокомпонентов.

А₆В₂. Увеличение широкополосности (быстродействия) всех элементов РЭА, интегральных микросхем и функциональных компонентов.

А₆В₃. Расширение областей применения опто- и акустоэлектронных приборов (оптических модемов, приборов на поверхностных акустических волнах, приборов с зарядовой связью (ПЗС) и др.).

А₆В₄. Повышение надежности, снижение стоимости электронных и микроэлектронных приборов и элементов.

Уровень С

Подцель А₁В₄

А₁В₄С₁. Совершенствование методов стабилизации частоты, разработка новых высокостабильных кварцевых резонаторов.

А₁В₄С₂. Разработка новых транзисторов, совершенствование модульных широкополосных усилителей мощности, разработка принципиально новых методов сложения мощности.

Подцель А₂В₄

А₂В₄С₁. Разработка высокоскоростных систем передачи цифровых сигналов.

А₂В₄С₂. Совершенствование адаптивных полиномиальных и гармонических корректоров временных и частотных характеристик каналов связи.

А₂В₄С₃. Разработка методов комплексной коррекции (оптимизация формы передаваемого сигнала и оптимальный радиоприем), применение методов адаптивного предсказания и корректирования сигналов.

Подцель А₄В₄

А₄В₄С₁. Разработка теоретико-информационных и теоретико-игровых методов синтеза оптимальных систем передачи информации.

А₄В₄С₂. Совершенствование методов формирования сложных сигналов, разработка новых принципов построения широкополосных систем связи с шумоподобными сигналами.

Как видно из рис. 1.2, построение дерева целей как главная составная часть целевой разработки научно-технической программы основано на последовательной детализации целей по линии: **главная цель** — **подцели первого уровня** — **подцели второго уровня** и т. д. Основой такой детализации является функционально-содержательный принцип, состоящий в том, что каждый целевой компонент разбивается на

целевые компоненты того же содержания, но только на более детальные.

В принципе дерево цели можно развернуть и по чисто производственному признаку, например, при нулевом уровне — повышение эффективности радиоэлектронного производства.

1.4.2. Предмет исследования

Предмет (тема) исследования во многом определяется целью, но к нему предъявляют и особые требования.

Во-первых, предмет исследования должен быть типичным, то есть характерным для сущности исследования.

Чтобы избежать ошибок, следует точнее раскрыть понятие типа, типичного.

Выделение из множества объектов типа, типичного предполагает выполнение работы, аналогичной классификации, являющейся основой многих исследований.

Тип или класс — понятие о группе из множества, определяемое общим существенным признаком (или совокупностью признаков), присущим только данному типу или классу. **Все классы множества составляют систему.** Поэтому перед определением типа, типичного надо вначале определить **состав признаков по их качеству**, затем, если потребуется, однородные группы (типы) разбить на подтипы. Если необходимо количественное определение типов и подтипов, эти данные берут в определенных пределах, не перекрывающих друг друга. Количественное разнообразие подтипов отображают при помощи **рандомизации (случайного отбора членов выборки)**. Например, ставится цель — найти оптимальные условия уборки кукурузы на силос. Берем типы кукурузы: сорт, чистая или с подсевом, богара или поливная; подтипы по густоте стеблестоя, диаметру стебля на уровне среза, высоте растений. Выбрав подходящий тип и подтип, поле (если оно слишком велико) делим на участки, нумеруем последние и из всех выбираем случайным образом (по таблице случайных чисел или по последним цифрам шестизначных тригонометрических таблиц) необходимое число.

При исследовании свойств почвы, например, выбирают почву, типичную для данного района.

Сравнительные испытания машин нельзя проводить, если одна машина новая, а другая изношена. Для исследования нет смысла брать явно устаревшие, снятые с производства машины, и т. д.

Во-вторых, предмет исследования должен быть строго определенным, нерасплывчатым. Глубина и тщательность разработки экспериментальной темы определяют качество исследования.

В-третьих, чтобы исследование могло быть выполнено, его тема должна соответствовать материальной базе.

В-четвертых, каждое исследование должно содержать экономическую часть.

В качестве примера ниже приводится перечень основных направлений научно-исследовательских работ в области машиностроения.

1. ***Исследование технологических процессов.*** Необходимо создавать и изучать новые, более совершенные процессы и приемы. Технологические процессы, которые нельзя заменить более совершенными, надо видоизменить в соответствии с естественными особенностями исследуемых процессов. Такая работа (уточнение технологии применительно к существующим условиям) может дать большой производственный эффект. Очень важно теоретическое и практическое установление полей допусков на качество работ.

2. ***Оценка конструкций машин, отдельных механизмов и рабочих узлов в разных условиях.*** Усовершенствование рабочих узлов и деталей. Определение качества работы машины или рабочих узлов, износов различных деталей и влияния их на качество и стоимость работы и установление предельных износов. Исследование кинематики и динамики конкретных механизмов, машин и машинных агрегатов, а также их энергетика. Определение надежности и экономических показателей машин и агрегатов. Поиски оптимальной надежности.

Первые два направления могут быть связаны с конструированием новых машин или усовершенствованием уже имеющихся. Если создание новой конструкции затруднительно, обоснование системы машин и требований к конструкциям может являться вполне достаточным результатом научного исследования. Основные цели исследований по первым двум направлениям — улучшить качество работы, повысить производительность труда и уменьшить затраты.

3. ***Изучение свойств материалов, используемых в машинах, и изменения этих свойств в различных условиях.*** Эта тематика в одинаковой степени важна и для конструирования машин и для рационального их использования.

4. ***Исследование технического обслуживания (эксплуатации и ремонта) машин.*** Основная цель этих исследований — поддержание технической надежности машин на высоком уровне, уменьшение простоев из-за неисправности и ремонта, а также снижение затрат на техническое обслуживание.

5. Изучение общих вопросов использования машин: эксплуатационных свойств машин; составления машинных агрегатов; режимов их работы и энергетики; нормирования, использования транспортных средств; нахождения оптимальной структуры машинного парка для данных типовых (модельных) условий; прогнозирования нужного количества машин новых типов, конструкций новых машин и устройств, необходимого количества ремонтов; запчастей и узлов обменного фонда и т. п.

Важны работы, связанные с увеличением скорости движения машинных агрегатов. В частности, к этой группе можно отнести: конструирование, улучшение и испытания устройств для скоростных машин, определение оптимальных скоростей и способов движения для конкретных условий; способы наилучшей подготовки агрегатов к работе на повышенных скоростях; механизация вспомогательных процессов и другие исследования, связанные с внедрением повышенных скоростей.

6. Исследование общетехнических вопросов: от исследования напряжений в деталях машин до применения автоматики и телемеханики. Здесь с экспериментальным исследованием тесно связано развитие различных теоретических обобщений.

В заключение отметим, что, если есть надлежащие средства и важность исследуемой проблемы, то сложность проблемных исследований не должно смущать молодых исследователей: «не боги горшки обжигают».

1.5. Рабочая гипотеза

Начинать исследования следует с разработки рабочей гипотезы. Для составления рабочей гипотезы следует тщательно изучить литературные данные (отечественные и зарубежные), все доступные отчеты о проведенных ранее аналогичных исследованиях.

Все эти материалы необходимо критически проанализировать, чтобы выяснить, что уже достигнуто и разработано, какие оригинальные направления и творческие замыслы развивались для решения задачи, какие есть неясности, противоречия и недоработки. Выявить, если это возможно, методические ошибки и недостатки предыдущих исследований, какие намечаются пути к открытию нового и развитию, улучшению старого. К анализируемым работам нельзя подходить предвзято, надо уметь видеть главное — *физический смысл явлений*.

Тщательно изучив все материалы, относящиеся к цели и предмету исследования, разрабатывают его рабочую гипотезу. Как уже

упоминалось, **рабочая гипотеза**—научное предположение о развитии явлений и их объяснение; это предположение не доказано, но в той или иной степени верно.

Таким образом, рабочая гипотеза имеет систематизирующее значение. Иногда ее целесообразно начать с **классификации** изучаемых сторон явления или факторов, его определяющих.

Вполне возможно, что перед исследователем откроется несколько путей решения, соответствующих различным вариантам рабочей гипотезы. Все варианты надо тщательно продумать и записать. Исследователь принимает тот вариант, который, по его мнению, дает наиболее быстрое и верное решение, и в ходе исследования не должен забывать, что данные опытов могут не соответствовать принятому варианту, но зато подтвердить жизненность другого. Реже представляется возможность вести исследование сразу по нескольким вариантам гипотезы.

Как минимум, рабочая гипотеза устанавливает факторы (причины, основные условия, движущие силы, иначе говоря — *измеримые переменные величины, воздействующие на объект исследования*). Рекомендуется вначале записать возможно более полный перечень всех факторов, которые предположительно могут действовать на развитие явления. Отсюда следует, что необходимо одновременно обсудить возможность придавать любому фактору требуемое в определенных границах значение (так называемая управляемость объекта). Следует не забывать при этом, что в некоторых случаях степень управляемости объекта связана с затратами на производство экспериментов, это весьма важная исходная часть методики; при неправильном определении факторов исследование может быть безрезультатным, так как пойдет не в том направлении, в каком развивается явление, и поэтому приведет к ложным выводам.

Как максимум, рабочая гипотеза дает более или менее полное предположительное объяснение всего процесса развития явления.

В принятой рабочей гипотезе необходимо выделить из всех уже известных наиболее важные и решающие причинные связи и взаимодействия и наметить возможные связи, взаимодействия, вероятное направление и ход развития. При всем этом рабочая гипотеза должна быть логически проста и во всех своих деталях проверяема экспериментально. Изложена она должна быть кратко и ясно. **Особое внимание следует обратить на точность понятий и терминов.**

Хотя часто словесное изложение гипотезы вполне возможно, в исследованиях такое изложение обычно дополняют графическим

изображением предполагаемых функциональных связей. В некоторых случаях, а именно там, где главные факторы и связи основное направление и ход развития явления не вызывают сомнений, развитие явления полезно представить в виде **математических моделей, в которых кривые развития явления заменены системой взаимосвязанных математических формул.**

Иногда для сокращения времени и средств экспериментальное исследование разделяют на исследования при помощи физических моделей и результирующее исследование принятых объектов. Эти особенности изложения рабочих гипотез полезно кратко рассмотреть.

В графиках прежде всего представляют зависимости результатов от изменения в принятых пределах основных факторов; далее могут быть развиты связи между факторами или другие зависимости, принятые в гипотезе. В простом случае несколько графиком представляют зависимости в координатах x, y . В более сложных случаях строятся пространственные графики (в координатах x, y, z), так что результат обычно будет представлен поверхностью. Наконец, если пространство закономерных взаимосвязей n -мерно, выбирают главные связи и по ним **строят ряд трехмерных графиков, так, чтобы каждый последующий график был связан с предыдущим одной из осей.**

Если какое-либо искомое является функцией ряда факторов, например $x=f(u, w, y, z)$, то при независимости этих факторов друг от друга легко построить графики $x=f_1(u, w)$; $x=f_2(y, z)$. В более сложных случаях (функция от функции, передаточная функция, взаимозависимые функции и т. п.) следует дать графики наиболее логично представляющие развитие явления. Например, если в только что приведенном четырехфакторном уравнении $w=f(y, z)$, то вначале строят пространственный график, а затем уже и график $x=f(u, w)$.

Все графики желательно, если имеются основания, давать с экстремумами и другими характерными точками (или частями поверхности). Такие графики не только показывают, что и в каких пределах следует измерять, но могут дать и принципиальное (не численное) решение задачи, и тем точнее, чем ближе к истине показанное на них развитие явления.

Весьма полезно предполагаемое развитие явления разработать в виде **математической модели.** Так же, как и при представлении рабочей гипотезы в виде графиков, здесь важно понимать развитие явления физически.

Под математической моделью явления понимается представление развития явления формулой или формулами.

В создании формул гипотезы наиболее важным действием является выбор или составление (на основании логических предпосылок) **структуры или типа формул**, в которых нашли бы себе отражение основные черты характера связи главных факторов.

Следует отметить, что одно и то же явление, один и тот же процесс можно описать при помощи различных математических моделей, пригодных для дальнейшей разработки.

В формулах целесообразно учитывать специфические свойства факторов (например, характер колебаний), наложить необходимые ограничения. Следует обдумать целесообразность аппроксимаций и линеаризации нелинейных связей. Во всяком случае математическая модель должна быть достаточно проста и *допускать (в соответствии с результатами опытов) изменение как структуры формулы, так и характера включенных величин и граничных условий*. В исследованиях математическая модель гипотезы часто представляется системой линейных дифференциальных уравнений, но вполне возможны и упрощенные структуры. В некоторых случаях математическую модель гипотезы полезно добавлять или разъяснять соответственно составленными графиками.

*Некоторым видоизменением метода математических моделей является метод **аналогий**, под которым понимается использование сходства признаков или отношений между исследуемым явлением и другим, имеющим разработанную и проверенную теорию.*

Если исследуемое явление достаточно сложно, самостоятельное составление математической модели может отнять слишком много времени. Естественно напрашивается вывод: взять уже разработанный и апробированный математический аппарат для внешне сходного, подобного процесса; в этом случае готовую математическую модель применяют по аналогии с исследуемой моделью. В таблице 1.1 указаны примеры некоторых аналогий.

При пользовании методом аналогий всегда следует учитывать, что внешнее сходство явлений может лишь заключать ту или иную степень вероятности сходства математических моделей. Например, применить теорию штампов к установлению законов резания лезвием было бы необдуманно, так как экстремумы здесь имеют прямо противоположные значения. Вместе с тем метод аналогий, логически анализируемый, представляет большую эвристическую ценность.

Таблица 1.1

Примеры аналогий

Объект исследования	Возможная математическая модель или группа моделей	Отрасль науки, в которой излагается модель подобного явления
Теория штампа	Топология (Римановы пространства)	Математика
Импульсные системы; кривизна четырехмерного пространства	Тензорный анализ	»
Взаимоотношения между системой и средой	Теория игр	»
Энергопередача; динамика множеств	Теплопередача	Учение о теплоте
Энергия связи частиц; качественные изменения и превращения динамических систем	Термодинамика	» » »
Схемы: накопление — разряд	Электростатика	Учение об электричестве
Волновые явления	Электродинамика	То же
Колебания конкретных машин и устройств	Теория колебаний	Механика
Точность движения; устойчивость комплексных упругих систем	Теория устойчивости движения	»
Пнеumo- и гидроизмерительные системы, процессы перемещения газов и жидкостей	Газовая и гидродинамика	»

Построенные любым образом математические модели гипотез развития явления сами по себе гипотетичны и, конечно, не дают окончательного изображения искомым закономерностей. Поэтому еще до окончательного формирования методики экспериментального исследования следует проверить математические модели, что легко выполнимо при помощи современных вычислительных машин. Обычные показатели для проверки — экстремальные величины (нуль, максимум, минимум, уход в бесконечность при принятых ограничениях), амплитуды и частоты колебаний.

Если обнаруживается несоответствие математической модели общей логической гипотезе развития явления, обсуждается внесение целесообразных поправок в модель.

Исследование рабочей гипотезы, представленной в любом виде, имеет для экспериментатора неопценимое организующее значение. Кроме того, в результате разработки и анализа гипотезы становится ясным: что, где, когда и как необходимо измерять в экспериментах.

Особняком стоит физическое моделирование, смысл которого заключается в том, что по результатам опытов на осуществленной модели судят о развитии явления в природе. По существу, построение рабочей гипотезы для таких исследований не отличается от изложенного ранее. Но обычно сами физические модели просты, количество действующих факторов уменьшено, и поэтому сущность рабочей гипотезы более проста.

Приведем пример (упрощенный, методический). ***Назовем его первым и разберем постепенно до конца.***

Пусть необходимо исследовать влияние затупления D лемехов на глубину вспашки h и сменную производительность W плужного агрегата.

Из литературы, отчетов по испытаниям плугов и производственного опыта выясняем, что по мере затупления лемехов тяговое сопротивление R плуга растет и он (при одной и той же регулировке) в среднем пашет мельче. Однако другие последствия затупления и характер отмеченных закономерностей остаются неясными.

Сделаем некоторые предположения. Конструкция лемеха ставит физический предел его затуплению, поэтому сопротивление плуга и средняя глубина пахоты в функции затупления лемеха должны иметь предел и в общем случае могут быть представлены в виде интегральных кривых (что наиболее вероятно). Но, помимо средней глубины вспашки, нас интересуют также отклонения O от средней. Они могут оставаться стабильными (O_2), увеличиваться (O_1) и уменьшаться (O_3) в функции затупления лемеха.

Все эти предположения составляют один вариант рабочей гипотезы. Другие предположения, основанные на том, что затупление лемеха не изменяет показателей его работы или улучшает их, отбрасываются как маловероятные.

На рисунке 1.3 представлен график, показывающий закономерности развития явления по принятой гипотезе.

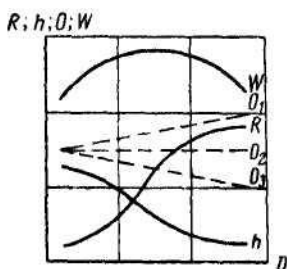


Рис. 1.3. Закономерности развития явлений по рабочей гипотезе (первый пример).

Для его построения величины аргумента откладываем по абсциссе.

В нашем примере по рабочей гипотезе следует измерить затупление лемехов, сопротивление плуга, глубину пахоты, производительность агрегата, и также прохронометрировать замену лемехов и сменное время.

Приведем еще один (*второй*), упрощенный для наглядности, пример. На этом примере *цель исследования* — установить влияние микронеровностей поля (борозд, гребней, глыб и др.) на колебания ножевого бруса жатки. Известно, что с увеличением высоты (или глубины) препятствий колебания остова машины и ножевого бруса растут; эти колебания зависят также от скорости движения, свойств почвы и эластичности шин жатки.

Учитывая все это, предполагаем, что величины абсолютных колебаний середины ножевого бруса жатки в функции величины препятствий и скорости движения будут изменяться так, как это показано на рисунке 1.4.

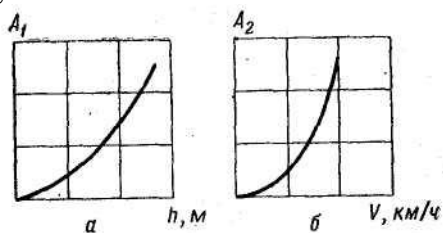


Рис. 1.4. Закономерности изменения величины A колебаний середины ножевого бруса жатки:

a — в функции высоты препятствий (микрорельефа) при неизменной скорости подхода к препятствию; b — в функции скорости движения при одной и той же высоте препятствий.

По этой гипотезе принимаем, что давление в шинах постоянно, что измерять следует величины колебаний середины ножевого бруса; геометрической оси машины, проходящей через центры подшипников колес и скорости; величины же неровностей поля (их характеристики для данного поля) измерить отдельно, например, профилированием.

В описанных схематических примерах рабочих гипотез мы выяснили, какие величины следует измерять в опытах, но рабочая гипотеза не устанавливает способов измерения этих величин.

1.6. Особенности исследования непрерывных процессов

В настоящее время очень редко приходится сталкиваться с необходимостью изучать обособленный технологический процесс. Чаще всего исследуемый объект представляет собой совокупность связанных процессов (говорят, что исследуемый объект обладает структурой). Очевидно, применение методов планирования эксперимента к таким объектам с одной стороны требует выработки новых рекомендаций и приемов, позволяющих решить поставленные задачи моделирования и оптимизации, с другой стороны — для реализации эксперимента на сложном технологическом процессе (будем называть его *системой*), требуется своя система исследования, представляющая собой объединение технологической установки, ЭВМ, исследователя и некоторых правил действий (или алгоритмов), обеспечивающих достижение поставленных целей. На рис. 1.5 представлена схема такой системы исследования.

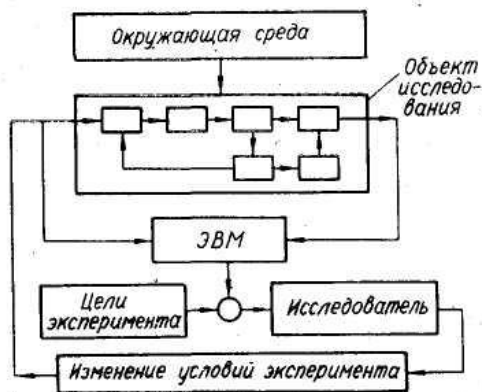


Рис. 1.5. Схема системы исследования сложного объекта.

Глобальную задачу исследования как целенаправленного процесса можно сформулировать так: *в соответствии с некоторым заданием, вызванным потребностью рынка в определенном продукте, необходимо получить информацию о способе производства этого продукта (совокупности процессов), достаточную для проектирования технологического комплекса заданной мощности.*

Решение глобальной задачи в значительной степени зависит от методов и средств, имеющихся в распоряжении исследователя. Будем считать, что исследователь имеет возможность проводить эксперименты на технологических установках определенного масштаба, пользоваться методами теории эксперимента (а значит, и вычислительной машиной). Исследователь ограничен временными, физическими и материальными ресурсами.

Очевидно, **решение глобальной задачи исследования** с системных позиций — *это решение сложной задачи оптимизации по многим критериям в условиях значительных ограничений и неопределенности.* В настоящее время эта задача формализована не в полной мере, тем не менее, она успешно решается исследователями с использованием проверенных практикой правил и принципов.

Здесь мы рассмотрим особенности применения методов планирования эксперимента к сложным технологическим системам на примере химической технологии, а также правил, объединенных некоторой логической схемой достижения поставленных целей.

1.6.1. Постановка задач исследования сложных процессов

Постановку задачи исследования сложных процессов осуществим на примере сложных объектов химической технологии.

При исследовании сложных объектов химической технологии встречается ограниченное число типов соединения процессов в системы. Принято **пять типов**: *последовательное, последовательно-обводное (с байпасом), параллельное, обратное (с рециклом) и перекрестное.*

Учитывая, что с точки зрения использования методов теории эксперимента и оптимизации системы с байпасом и последовательным соединением, а также системы с рециклом и перекрестные почти идентичны, ниже рассматриваются только три типа соединения процессов — последовательное, параллельное и с рециклом (рис. 1.6).

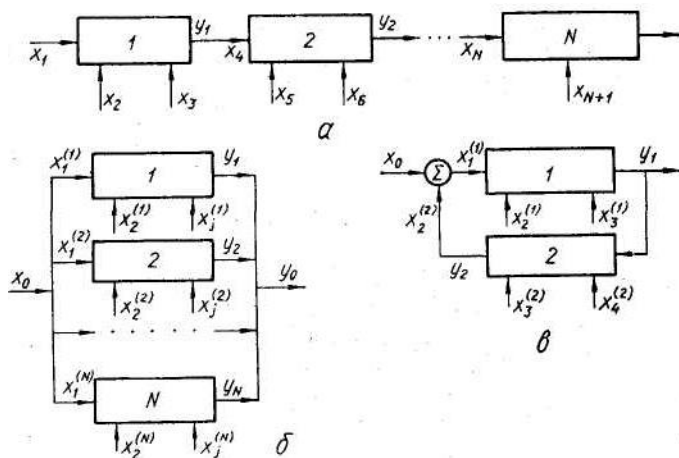


Рис. 1.6. Типы соединения процессов в системах исследования:
 а — последовательное; б — параллельное; в — соединение с рециклом.

Исследование системы с последовательным соединением процессов. *Исходные данные.* Имеется система, процессы которой соединены последовательно (см. рис. 1.6, а) технологическими потоками, т. е. система обладает структурой, которую можно выразить условиями:

$$\begin{aligned} y_1 &= x_4; \\ y_2 &= x_7; \\ &\dots \end{aligned}$$

в общем виде для i - и k -го процессов

$$y_k = x_i. \quad (1.1)$$

Необходимо найти математическую модель системы, используя методы планирования эксперимента, и решить задачу оптимизации системы. Математическая модель ищется в виде регрессионного уравнения

$$\hat{y}_N = f(x_1, x_2, \dots, x_{N+1}, b_0, b_1, \dots, b_{N+1}), \quad (1.2)$$

а задача оптимизации

$$R = \text{extr}_x y_N,$$

где R — экстремальное значение функции цели.

План эксперимента. К каждому процессу, начиная с первого, применяется соответствующий план эксперимента и определяются модели

$$\begin{aligned} y_1 &= f_1(x_1, x_2, x_3, b_j^{(1)}); \\ y_2 &= f_2(x_4, x_5, x_6, b_j^{(2)}); \\ &\dots \\ y_N &= f_N(x_N, x_{N+1}, b_j^{(N)}). \end{aligned} \tag{1.4}$$

При этом найденное значение выходной переменной y_1 (при постановке планируемого эксперимента на процессе 1) является фактором x_4 (при исследовании процесса 2 (см. рис. 1.6, а). Желательно при этом решить автономную задачу оптимизации для процесса 1, т. е. найти

$$\text{extr}_{x_1, x_2, x_3} y_1 = R_1, \tag{1.5}$$

и в качестве нулевого уровня фактора x_4 использовать

$$R_1 = \text{extr } y_1.$$

Решение задачи. Если не наметится противоречий в физико-химических связях уравнения

$$\text{extr } y_h = x_i, \tag{1.6}$$

то, зафиксировав фактор $x_N = \text{extr } y_{N-1}$ можно решить глобальную задачу оптимизации (1.3). Для этого достаточно найти

$$\text{extr } y_N = R_N = R \text{ при } x_N = \text{extr } y_{N-1} = \text{const}.$$

Если физико-химические ограничения не позволяют использовать условие (1.6), то для решения задачи (1.3) целесообразно использовать принцип динамического программирования: задается y_N и по модели

$$y_N = f(x_N, x_{N+1}, b_j^{(N)}) \text{ определяются соответствующие } x_N \text{ и } x_{N-1}.$$

Ранее задается $x_N = y_{N-1}$ и решается задача поиска соответствующих факторов для блока $(N-1)$. Расчетная процедура поиска оптимальных факторов (наверное, обратная рассмотренной выше прямой процедуре) продолжается до тех пор, пока не будут найдены x_1, x_2, x_3 . Полученные результаты проверяются на экспериментальной установке и, если расхождения невелики, принимается найденный технологический режим.

Замечание 1. Если в процессе оптимизации технологического режима на установке по каким-либо факторам будут достигнуты граничные условия, то они могут быть зафиксированы, и таким

образом исключены из процедуры планируемого эксперимента и поиска оптимальных значений выходных переменных.

Замечание 2. Если процессы в системе или сама система характеризуется больше чем одной переменной, то задача усложняется и переходит в область методов многокритериальной оптимизации.

Исследование системы с параллельным соединением процессов. Исследуется система, процессы в которой соединены параллельно общими технологическими потоками (см. рис. 1.6, б). Структура системы может характеризоваться условиями

$$x_0 = \sum_{i=1}^N x_i^{(i)}; \tag{1.7}$$

$$y_0 = \sum_{i=1}^N y_i. \tag{1.8}$$

Необходимо найти математическую модель системы, используя методы планирования эксперимента, и решить задачу оптимизации системы. Математическая модель ищется в виде

$$y_i = f_i(x_1^{(i)}, x_2^{(i)}, \dots, x_j^{(i)}, \dots, b_j^{(i)}), \tag{1.9}$$

где $i = 1, 2, \dots, N$ — номер процесса.

Задачу оптимизации можно записать так:

$$R = \text{extr}_x y_0 = \sum_i R_i = \text{extr}_x \sum_i y_i(x_1^{(i)}, x_j^{(i)}, \dots, b_j^{(i)}). \tag{1.10}$$

Если выполняется условие (1.7) и возможно измерение, а также управление факторами $x_1^{(1)}, x_1^{(2)}, \dots, x_1^{(N)}$, то задача получения математической модели системы в виде (1.9) методами планирования эксперимента распадается на отдельные процедуры по числу блоков в исследуемой системе. Таким образом, математическая модель системы, состоящая из параллельно соединенных процессов, представляет систему линейных и нелинейных регрессионных уравнений

$$\begin{aligned} y_1 &= f_1(x_1^{(1)}, x_2^{(1)}, \dots, x_j^{(1)}, \dots, b_j^{(1)}); \\ y_2 &= f_2(x_1^{(2)}, x_2^{(2)}, \dots, x_j^{(2)}, \dots, b_j^{(2)}); \\ &\dots \\ y_N &= f_N(x_1^{(N)}, x_2^{(N)}, \dots, x_j^{(N)}, \dots, b_j^{(N)}). \end{aligned} \tag{1.11}$$

Решение задачи оптимизации. При выполнении условия (1.8) задача оптимизации системы также распадается на автономные задачи поиска оптимума по каждому процессу:

$$R_i = \text{extr } y_i(x_1^{(i)}, x_2^{(i)}, \dots, b_j^{(i)}), \quad (1.12)$$

а далее оптимум системы находится в соответствии с условием (1.10).

Условие (1.8), как правило, выполняется для сложных объектов, у которых функция цели y_i совпадает с выходной переменной, характеризующей технологический режим (например, содержание некоторого продукта A в выходных потоках технологической системы). В тех случаях, когда y_i является показателем эффективности работы системы (наджность, устойчивость, рентабельность, стоимость и др.), условие (1.8) не выполняется, и тогда используют взвешенные оценки ρ_i для y_i , получаемые экспертизами.

Исследование систем с рециклом.

Значительно сложнее решаются задачи моделирования и оптимизации, когда в схеме имеется один или несколько рециклов. Структуру такой системы (см. рис. 1.6, в) можно выразить уравнениями:

$$x_1^{(1)} = x_0 + x_2^{(2)}; \quad (1.13)$$

$$y_2 = x_2^{(2)}. \quad (1.14)$$

Ставится задача: найти математическую модель системы методами теории эксперимента и определить оптимум y_1 :

$$R = \text{extr}_x y_1 \quad (1.15)$$

при

$$y_1 = f_1(x_0, x_2^{(1)}, x_3^{(1)}, b_j^{(1)}); \quad (1.16)$$

$$y_2 = f_2(x_3^{(2)}, x_4^{(2)}, b_j^{(2)}). \quad (1.17)$$

Замечание. Возможен рецикл вида $y_1=y_2$, т. е. выходной поток объекта 1, характеризующийся переменной y_1 направляется на вход того же объекта.

При наличии рецикла использование планов эксперимента для получения регрессионных моделей невозможно вследствие коррелированности факторов между собой (x_0 коррелировано с $x_2^{(1)}$ и др.). Напомним, что одной из предпосылок регрессионного анализа является отсутствие корреляции между факторами. При исследовании системы с рециклом возможны два случая.

1. Если существует возможность физически «разорвать» рецикл (между y_2 и $x_2^{(2)}$ поместить, например, емкость, накапливающую продукты выходного потока объекта 2), то моделирование сводится к соответствующей задаче последовательного соединения объектов в систему. Например для блока 1 (см. рис. 1.6, в) факторами будут $x_1^{(1)} = x_0 + x_2^{(2)}$ при $x_2^{(2)} = \text{const}$, $x_2^{(1)}$, $x_3^{(1)}$, а выходной переменной y_1 .

2. Если рецикл «разорвать» нельзя, то моделировать процессы в системе следует на основе физико-химических закономерностей. Идентификацию системы и оценку ее адекватности можно осуществить с использованием полиномов различных порядков. Для этого необходимо иметь выборку экспериментальных данных.

Решение задачи оптимизации. В первом случае, когда существует возможность «разрыва» рецикла, задача оптимизации системы решается так же, как и при последовательном соединении объектов. Находится

$$R_1 = \underset{x^{(1)}}{\text{extr}} y_1 \tag{1.18}$$

при условии, что модель в виде (1.16) получена. Далее определяется

$$R_2 = \underset{x^{(2)}}{\text{extr}} y_2. \tag{1.19}$$

«Стыковку» R_2 (оптимальное значение для y_2) и $x_2^{(2)}$ можно рассчитать, используя, например, итерацию.

Задача оптимизации системы по физико-химическим моделям рассчитывается с использованием целого арсенала поисковых методов.

1.6.2. Основные цели и задачи процесса исследования технологических систем

Как уже упоминалось ранее, процесс исследования технологии может быть представлен как некоторая система, предназначенная для достижения определенных целей. Все операции по исследованию технологических систем обычно связаны некоторой логической схемой «добывания» информации и использования ее для решения поставленных задач. В этой логической схеме исследования, методы планирования эксперимента позволяют решить отдельные задачи и задачу исследования в целом. В этом разделе рассмотрены основные принципы решения задач исследования, выделены цели логической схемы исследования и намечены пути использования системы алгоритмов теории эксперимента в существующем процессе исследования технологических систем.

1.6.2.1. Основные принципы исследования технологических систем.

Практикой исследования накоплено много принципов, реализация которых в значительной мере способствует эффективности проводимых экспериментов. Ниже рассмотрены основные из них. Изложенные принципы не ранжировались по их значимости.

Процесс исследования в целом можно рассматривать как совокупность операций над объектом исследования.

Объект исследования (лабораторная установка определенного масштаба), исследователь и ЭВМ с соответствующим математическим и программным обеспечением объединены в систему. Перед системой исследования поставлена цель. Вход в систему — существующая вне ее информация о свойствах продукта и способах его получения (возможно по аналогии с другими продуктами); выход — данные для проектирования предприятия определенной мощности.

Принцип 1. Процесс исследования — это анализ физических и химических явлений на глубину, достаточную для получения информации и передачи ее в систему проектирования.

Логическая схема исследования (ЛСИ) строится так, чтобы последовательно, шаг за шагом, увеличить информацию об объекте исследования до уровня, достаточного для проектирования технологического комплекса. ЛСИ предполагает также организацию последовательных приближений и повторных циклов исследований на отдельных этапах.

Принцип 2. При решении задач исследования всегда существует неопределенность, преодолеваемая исследователем (он принимает одно из альтернативных решений).

Принятие решений — это «стыковочный узел» этапов исследования. Именно здесь можно учесть неформализуемые условия — ресурсы творческой активности, прогнозы специалистов и т. п.

Принцип 3. Логическая схема исследования имеет глобальную цель (о ней говорилось ранее) и подцели, достижение которых позволяет решить задачу в целом. На каждом этапе, при достижении подцели, исследователь пользуется определенной информацией, полученной на предыдущем этапе или имеющейся вне системы. На всех этапах исследования можно пользоваться информационным обеспечением (ИО) данной подсистемы или ИО более высокого уровня..

Принцип 4. При решении задач в рамках системы исследования удобно и оправдано пользоваться математическими моделями физических и химических явлений. Однако, следует отметить, что на

протяжении исследования роль математических моделей меняется. Если обрабатывается технологический режим на лабораторной установке, то математическая модель используется как средство достижения оптимального технологического режима (например, максимального выхода продукта). Если «нарабатывается» информация для проектирования, то математическая модель является целью процесса исследования, и эта цель достигается некоторым оптимальным путем (минимизируются, например, ресурсы для достижения цели).

Принцип 5. В процессе исследования широко используется принцип декомпозиции технологической системы на технологические узлы. Логика построения технологических комплексов предлагает **четыре типа таких узлов**: подготовки сырья *A*, химического превращения *B*, разделения продуктов реакции *C* и переработки и обезвреживания отходов производства *D*. Декомпозиция ABCD-комплексов на составляющие узлы позволяет эффективно проводить новые исследования или пользоваться известной информацией с уточнением параметров по результатам эксперимента.

Принцип 6. В процессе исследования соблюдается принцип главного технологического узла (чаще всего это химический процесс). От степени изученности главного технологического узла в значительной степени зависят дальнейшие проектные решения. В процессе исследования соблюдается также принцип главного руководителя исследований, принимающего основные решения в логической схеме исследования.

Принцип 7. В процессе исследования существует иерархия принимаемых решений (например, принятие решений на уровне руководителя исследований и технического Совета).

Принцип 8. Процесс исследования имеет свой документооборот и свои способы его организации, в том числе и способы связи с процессом проектирования.

Перечисленные принципы являются основными в процессе исследования. Они сохраняются и при автоматизированном эксперименте, хотя организация всего процесса исследования, использование и переработка информации, безусловно, меняются.

Логические схемы исследования (ЛСИ). Существует множество логических схем исследования. Их разнообразие обусловлено возможностями современной науки, ограничениями, которые накладывают исследуемые процессы, и конкретными условиями существования данного исследовательского подразделения. Самым сильным ограничением, накладываемым на технологические процессы,

является неадекватность масштабного перехода от изучения явления на условном микроуровне к изучению условий на макроуровне. Информация, полученная на микроуровне, не может быть полностью использована на макроуровне. Приходится организовывать многоступенчатую систему исследования, причем шаг ступени нельзя заранее определить точно, хотя такие предложения есть.

Ниже рассмотрены *три ЛСИ* — две полярные по структуре и одна компромиссная. В методических целях структура рассматриваемых ЛСИ значительно упрощена.

Многоступенчатая ЛСИ. Эта логическая схема более всего отражает трудности, с которыми приходится сталкиваться, когда учитывается неадекватность масштабного перехода. Многоступенчатость вызвана стремлением учесть неадекватность масштабного перехода. Заранее предполагается, что есть такой шаг масштабного перехода, в пределах которого неадекватность сводится к минимуму или практически ею можно пренебречь. К сожалению, выбор величины этого шага ни теоретически, ни экономически не доказан. Существует практическая многоступенчатая ЛСИ, ее общий (сокращенный) вид приведен на рис. 1.7.

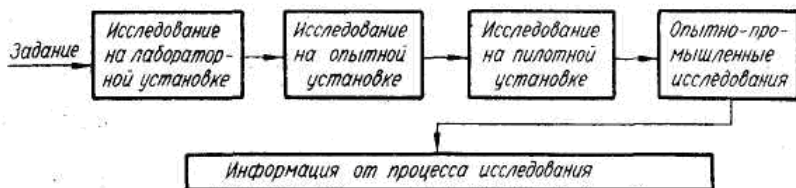


Рис. 1.7. Многоступенчатая логическая схема исследования.

Число ступеней может быть большим или меньшим в зависимости от сложности исследуемых процессов и заданной глубины исследования.

В результате деятельности такой ЛСИ проектировщики получали значительную информацию — условия технологического режима, достаточно близкие к оптимальным, и апробированную структуру технологического комплекса (ведь опытно-промышленная установка по структуре почти идентична промышленной). Но такая ЛСИ дорогостоящая, и главный ее недостаток — длительность исследования. Современные условия проектирования требуют сокращения непроизводительных этапов «жизненного» цикла технологического производства.

Многоступенчатая ЛСИ не потеряла актуальности. Она используется там, где требуются фундаментальные исследования процессов (или производств), использование которых прогнозируется на долгие годы. Эту систему можно считать крайним полюсом в различных логических схемах.

Одноступенчатая ЛСИ. Другим полюсом ЛСИ является схема, изображенная на рис. 1.8. Ее лаконичность обусловлена применением метода математического моделирования.



Рис. 1.8. Одноступенчатая логическая схема исследования.

Принципиально возможно провести исследования на лабораторной установке и получить полную математическую модель системы, адекватную масштабному переходу, а затем предоставить эту информацию для проектирования предприятий любой мощности. Однако некоторые специалисты считают, что расходование ресурсов в многоступенчатой и одноступенчатой схемах примерно одинаково (сравнительный анализ этих ЛСИ не проводился). Это связано с тем, что реализация одноступенчатой ЛСИ требует очень тонких и дорогостоящих исследований в рамках системы автоматизированного эксперимента. Тем не менее, ожидаемое сокращение времени исследования позволяет считать, что в будущее за использованием одноступенчатых логических схем исследования.

Компромиссная ЛСИ. Для большинства технологических исследований тех производств, которые часто меняются, значительные затраты на исследования не оправданы. Поэтому предлагаются и существуют различные компромиссные схемы исследования. Компромисс находят в основном между глубиной исследования и временем выдачи результатов. В рассматриваемой здесь компромиссной ЛСИ это достигается за счет использования ряда узлов *A, B, C, D* различной производительности.

Синтез технологического комплекса осуществляется условно за счет емкостей накопления продукта. Это дает возможность не строить опытной установки, которая, как известно, должна быть согласована по производительности. Компромиссная принципиальная схема исследования представлена на рис. 1.9.

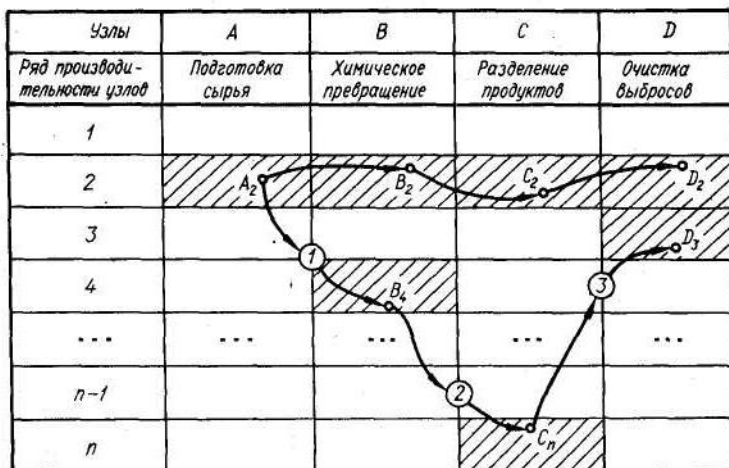


Рис. 1.9. Компромиссная принципиальная логическая схема исследования:

$A_2B_2C_2D_2$ — установка, согласованная по производительности узлов (соединение в схему возможно); $A_2B_4C_nD_2$ — модельная установка, не согласованная по производительности узлов (соединение в схему невозможно); \bigcirc — условная накопительная емкость.

Преимущество компромиссного подхода — быстрота получения информации за счет использования ряда узлов A, B, C, D различной производительности в модельном цехе. На рис. 1.9 пример такого соединения — $A_2B_4C_nD_3$: подготовка сырья в узле A производительностью 1, накопление сырья в емкости 1, химическое превращение в узле B производительностью 4, накопление продукта в емкости 2, разделение продукта на ректификационной колонне C производительностью n , накопление продуктов в емкости 3, очистка выбросов в узле D производительностью 3. В сочетании с принципом главного узла (этому узлу уделяется особое внимание, например, он может быть создан заново) такая схема имеет высокую эффективность. Она особенно хороша, когда есть общий вид математических моделей и необходимо уточнить их параметры (здесь резко возрастает роль **методов идентификации**).

Недостаток модельного компромиссного подхода — отсутствие сведений о технологическом режиме замкнутой схемы (аппараты различной производительности нельзя соединить в схему иначе как последовательно через емкости). Действующие рециклы возможны

только в схеме с аппаратами примерно равной производительности. Таким образом, материальные и тепловые балансы технологической схемы с рециклами (замкнутой схемы) можно рассчитать только на этапе проектирования.

Более полная, чем предыдущие ЛСИ, компромиссная логическая схема изображена на рис. 1.10.

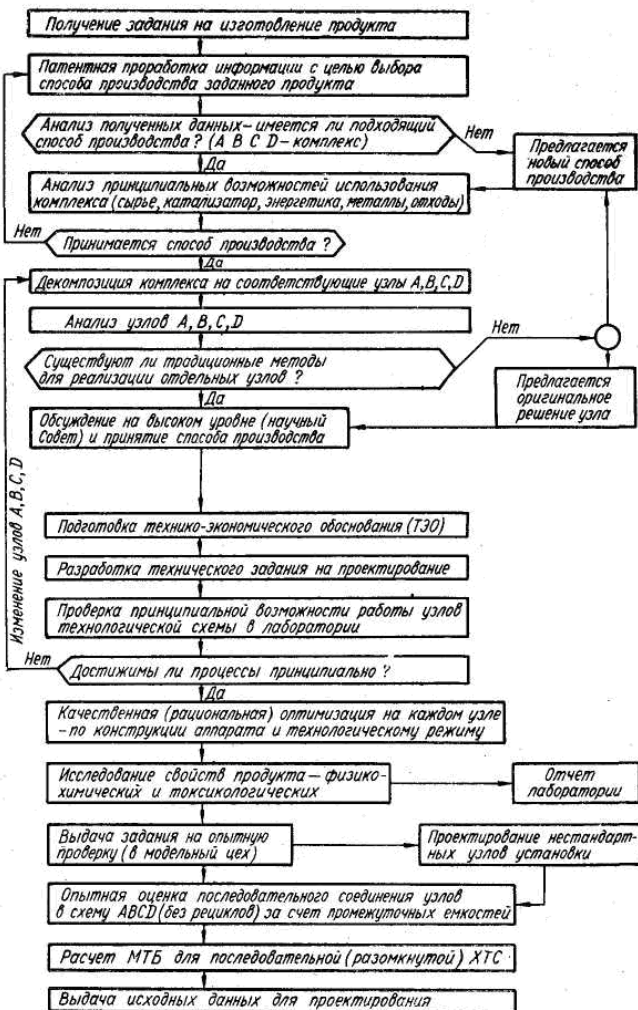


Рис. 1.10. Компромиссная логическая схема исследования.

Временная и объектная декомпозиция процесса исследования.

Выше упоминалось, что процесс анализа объекта исследования строится иерархически и главная его цель — получение информации, достаточной для проектирования. Также известно, что процесс исследования является итерационным — если на последующем этапе обнаружится недостаток информации или ее ошибочность, то исследователь возвращается на предыдущий этап и пополняет или исправляет информацию. Для организации этого итерационного процесса необходимо провести **декомпозицию** как по времени (этапам исследования), так и по объекту исследования (решение этой задачи в общем определяет логическую структуру системы автоматизированного эксперимента).

Выше рассматривалась компромиссная ЛСИ. В этой ЛСИ декомпозиция объекта исследования не идет далее узлов технологической схемы. Очевидно, что единых рецептов глубины декомпозиции предложить нельзя. Однако успешная работа системы исследования, включая и ЭВМ, зависит от тщательно аргументированной декомпозиции процесса исследования по времени и по объекту.

Предлагается по времени декомпонировать объект исследования на два этапа: *архитектурный А* и *функциональный Ф*.

Архитектурный этап предполагает *анализ структуры подсистем каждого уровня декомпозиции объекта исследования*.

Функциональный — *анализ функционирования подсистем некоторого уровня* (или работоспособности подсистем определенной структуры).

По объекту исследования предлагается выделить **четыре уровня декомпозиции**.

Первый уровень, наиболее общий — это уровень объекта (О) исследования в целом. На этом уровне анализируются способ производства заданного продукта, связи его с окружающей средой (рынком, отраслями и т. д.); возможности функционирования производства (изменения, вносимые в окружающую среду выпуском продукта), анализируются сырье, катализаторы, энергетика и т. д.

Второй уровень — это уровень анализа узлов *А, В, С, D* технологического комплекса (ТК), их связей с узлами промышленного здания и системой управления.

Третий уровень — это уровень анализа законов функционирования (ЗФ) процессов в узлах технологического комплекса (технологических аппаратах) в связях этих узлов друг с другом (трубопроводах) и системой управления.

Четвертый уровень связан с анализом элементарных законов функционирования в процессах и анализом существующих между явлениями связей. Выделение этого уровня сулит исследователям значительные выгоды, особенно при использовании ЭВМ.

Временная и объектная декомпозиция процесса исследования представлена в табл. 1.2.

Таблица 1.2.

Декомпозиция этапов и объектов исследования

T^* N	А	Ф
О	Анализ связей ТК с окружающей средой (рынок, другими отраслями; предприятием, где будет спроектирован комплекс)	Анализ функционирования ТК (способ производства продукта; анализ сырья, катализатора, энергетике, материала аппаратов, отходов; анализ динамики экономических отношений при сбыте продукта)
ТК	Анализ связей между узлами ТК; узлами и зданием; комплексом и системой управления	Анализ функционирования узлов А, В, С, D; узлов здания; открытых площадок, емкостей накопления продуктов, системы управления
ЗФ	Анализ законов функционирования: связей между аппаратами узлов, транспортными связями, управленческих связей (возможна лабораторная проверка)	Анализ функционирования процессов в аппаратах узлов (лабораторный масштабный уровень); исследование свойств продукта и свойств информационных процессов
ЭЗФ	Анализ элементарных законов функционирования — связей между элементарными процессами	Анализ функционирования элементарных процессов с лабораторной проверкой

Математические модели

T^* — этапы исследования; N — уровни декомпозиции объекта исследования.

Следует отметить, что каждый из предложенных уровней можно декомпозировать еще на несколько. Количество уровней, в конечном счете, определяется конкретным объектом.

На рис. 1.11 также показано, что процесс исследования всегда заканчивается математическими моделями явлений и связей. Если

обозначить индексом d уровни таблицы ($d=1, 2, 3, 4$), а i — этапы ($i=1,2$), то обозначение ячейки будет d_i , и тогда можно утверждать, что система полученных моделей $M\{M^{d_i}\}$ должна соответствовать матрице декомпозиции.

Формальное описание процесса исследования. Основная задача формального описания процесса исследования — получить такую логическую схему, которая позволила бы формальными средствами до окончания всего исследования произвести оценку разрешимости задачи вообще и установить, какие ресурсы понадобятся для достижения поставленной цели (получения математических моделей). Для формализации любой сложной системы используются идеи системного подхода и языки теории множеств, теории графов и, конечно, ЭВМ.

Прежде всего выделим ячейку исследования J — формализованное понятие, включающее ряд элементов и процедур: исходные данные A_j ; ограничение C_j ; математическую модель M_j (для достижения цели); решающую процедуру T_j ; оценку результата исследования K_j ; результат исследования (решение) R_j .

Эту явную часть ячейки исследования удобно записывать в виде множеств $\{A_j, C_j, M_j, T_j, K_j, R_j\}$.

Пояснение 1. Следует отличать математическую модель M_j ячейки J , которая предназначена для получения только результата R_j , от математической модели как цели всего процесса исследования.

Пояснение 2. Процедура T только в общем случае является решающей (множество функций цели, значительная неопределенность), а в частном случае — это может быть одним из случаев оптимизации. Например, в ячейке $\{ЗФ, Ф\}$ декомпозиционной матрицы обычно возникает задача поиска максимального выхода продукта на лабораторной установке, который может быть найден симплекс-методом по регрессионной модели второго порядка, полученной экспериментально.

Явной части ячейки исследования всегда сопутствует ее неявная часть J^1 : ресурсы исследования Z_j ; затраты на исследование L_j ; цели функционирования G_j .

Неявную часть ЯИ можно записать в виде множества $\{Z_j, L_j, G_j\}$. Эта оценка может быть прогностичной (до операций по достижению цели) и итоговой (после проведения исследований).

В каждой ячейке исследования формируются операторы или функциональные отношения F , отображающие одно множество в другое. Для перехода от $(J-1)$ -й к J -й и далее к $(J+1)$ -й ячейке можно записать

$$F_1 : R_{j-1} \longrightarrow \{A_j, C_j\}; \quad (1.20)$$

$$F_2 : R_j \longrightarrow \{A_{j+1}, C_{j+1}\}; \quad (1.21)$$

$$F_2^1 : \{Z_j, L_j, G_j\} \longrightarrow R_j. \quad (1.22)$$

Эта запись читается так. Функциональная система F_1 является отражением результата исследования предыдущей ($J-1$)-й ячейки во множество исходных данных $\{A_j, C_j\}$ j -й ячейки. Функциональная система F_2 является отражением результата исследования R_j , являющегося следствием $(A_j, C_j, M_j, T_j, K_j)$, во множество исходных данных $\{A_{j+1}, C_{j+1}\}$ для $(J+1)$ -й ячейки. На результат исследования влияет функциональная система F_2^1 , представляющая собой отражение множества оценок K_j , зависящего от множества $\{Z_j, L_j, G_j\}$, во множество результатов исследования R_j .

Функциональные системы F_1, F_2, \dots, F_N реализуются с помощью натурального эксперимента (как правило, на лабораторной установке, хотя в начальной стадии не исключены и экспертные оценки). Таким образом, исходные данные для каждой последующей системы «добываются» натурным экспериментом (в отличие от процесса проектирования, где они «добываются» математическим экспериментом на модели). Натурный эксперимент дает результаты, отмеченные неопределенностью масштабного перехода, а математический эксперимент — отмеченные неопределенностью прогноза.

Если считать декомпозиционную матрицу основой для создания ячеек исследования, то в целом исследование можно представить как некоторый итерационный процесс. Например, последовательность

$$F_1 : R^O, \Lambda \longrightarrow \{A^O, \Phi, CO, \Phi\}; \quad (1.23)$$

$$F_2 : R^{TK, \Phi} \longrightarrow \{ATK, \Phi, CTK, \Phi\}; \quad (1.24)$$

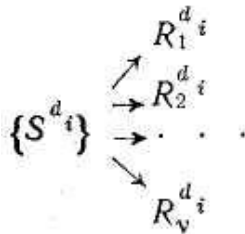
$$F_3 : R^{TK, \Phi} \longrightarrow \{ATK, \Lambda, CTK, \Lambda\}, \quad (1.25)$$

предполагает переход от анализа функционирования узлов системы к анализу ее связей и наоборот.

Двумерную систему задач исследования можно представить таблицей $\{S^{d_i}\}$, где $d = O, TK, 3\Phi, ЭЗ\Phi$; $i = A, \Phi$. Тогда можно предложить две идеализированные системы исследования: $\{S^{d_1}\} \rightarrow \{S^{d_2}\}$ — последовательную, декомпозированную только по этапам; $\{S^{1i}\} \rightarrow \{S^{2i}\}$ - параллельную, декомпозированную только по

объекту.

Системные задачи исследования не разрешимы ни по одной из этих идеализированных систем вследствие недостатка информации на предыдущем этапе (более глубокое исследование всегда таит неопределенность), а также вследствие неоднозначности решения при использовании решающей процедуры (альтернативные решения $v=1, 2, \dots$):



Возможно последовательно-параллельное осуществление процесса исследования:

1. $O \{A, \Phi\} \rightarrow TK \{A, \Phi\}$ — возможно технико-экономическое обоснование будущего ХТК;
2. $3\Phi \{A\} \rightarrow 3\Phi \{\Phi\}$ — возможно получение упрощенных моделей;
3. $\text{ЭЗФ} \{A\} \rightarrow 3\Phi \{\Phi\}$ — возможно получение полных математических моделей.

Изложенное выше позволяет говорить о целесообразности выделения двух систем в исследовании:

- а) процедурной, связанной с получением информации (фактически она изложена в матрице декомпозиции);
- б) целевой, связанной с получением информации на лабораторной установке.

Последнюю лучше всего представить в виде дерева декомпозиции объекта исследования (рис. 1.11).

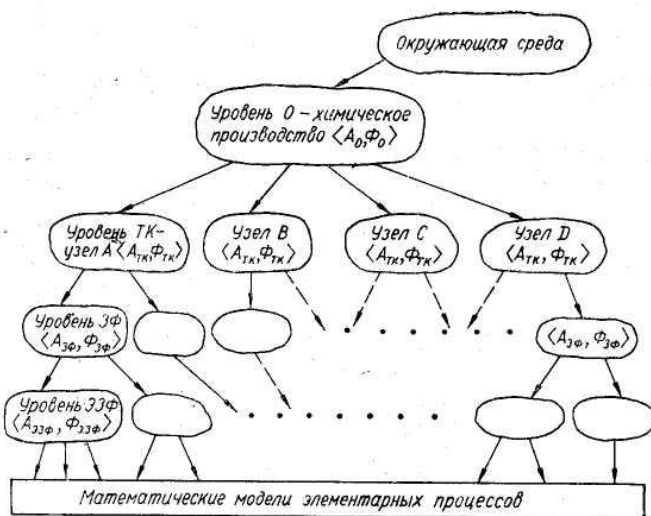


Рис. 1.11. Дерево декомпозиции объекта исследования.

Проведенный анализ позволяет также уточнить общую цель исследования. Следует выделить две разновидности целей:

Цель I, характерная для научного фундаментального познания,— это создание математических моделей без учета схемы производства, которая будет затем синтезироваться. Для этой цели характерно отсутствие итераций в исследовании.

Цель II, характерная для прикладных исследований,— это создание математических моделей для процессов того производства, которое будет создано на новом масштабном уровне. Для ЛСИ здесь характерны итерационные процессы (уточняются состав процессов, законы их функционирования и архитектура).

Выделение основных целей процесса исследования.

Предложенные выше уровни декомпозиции как объекта исследования, так и этапов его исследования можно разделить еще на несколько подуровней. Количество уровней, как уже указывалось, определяется стоящим перед исследователем конкретным объектом. Однако цели исследования в технологических процессах достаточно четко определены и могут быть выделены и отнесены к некоторому инварианту (неизменной части процесса исследования). Одновременно рассматриваются методы (или алгоритмы), позволяющие достичь выделенных целей.

Цель I — сбор априорной информации о методе производства заданного продукта, условиях проведения технологических процессов,

основных показателях технологии в целом и ее узлов.

Эта цель достигается без лабораторного натурального эксперимента. Исследователи проводят литературный и патентный обзор, т. е. анализируют имеющиеся в их распоряжении данные. Оценивают существующие или аналогичные способы производства заданного продукта, анализируют узлы схемы. Методы исследования — обычно экспертные оценки, в частности, метод ранговой корреляции (рис. 1.12, а).

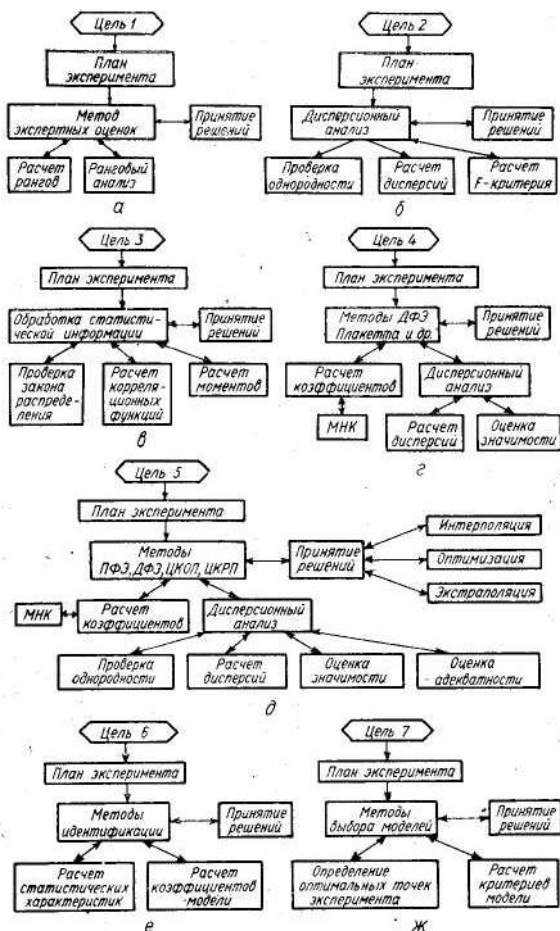


Рис. 1.12. Схемы реализации целей в процессе исследования технологических систем.

Системные элементы ячейки исследования для достижения цели представлены в табл. 1.3.

Таблица 1.3.

Системные элементы достижения цели 1

Элементы ЯИ ¹	Содержание элементов
A_1, C_1 — исходные данные и ограничения	Потребности народного хозяйства в продукте и промышленные возможности его получения
M_1 — математическая модель	Модель экспертных оценок, план постановки экспертиз и обработка результатов (усредненная оценка экспертов)
T_1 — решающая процедура	При согласованности оценок экспертов принимается их усредненная оценка или используется мнение главного руководителя исследований
R_1 — результаты исследования	Способ производства, узлы производства, список возможного сырья и катализаторов; энергетические решения, материалы аппаратов, использование отходов
K_1 — оценка результатов исследования	Используются технико-экономические показатели объекта исследования и процесса исследования с точки зрения ресурсов, затрат и достижения цели

¹ ЯИ — ячейка исследования.

Цель 2 — выбор факторов технологического комплекса. Прежде всего это выбор сырья и катализаторов, определяющих выход заданного продукта, и затем выбор режимных переменных технологического комплекса.

Цель достигается постановкой экспериментов на лабораторной установке. Задача решается методами дисперсионного анализа или с применением планирования по схеме латинских квадратов. Схема реализации этой цели изображена на рис. 1.12, б, а системные элементы процесса ее достижения приведены в табл. 1.4. Заметим, что режимные факторы выбираются исследователем по аналогии с другими производствами, возможен анализ экспертных оценок.

Таблица 1.4.

Системные элементы достижения цели 2

Элементы ЯИ	Содержание элементов
A_2, C_2	Задан способ производства продукта и составляющие его технологические узлы; есть перечень типов сырья и катализаторов; имеются режимные переменные
M_2	Модели дисперсионного анализа, различные планы его реализации
T_2	Соответствие статистическому критерию Фишера с учетом мнения исследователя
R_2	Сырье и катализаторы, дающие наибольший выход продукта, и режимные переменные
K_2	Оценка технико-экономических показателей, полученных результатов — сырья и катализаторов, а также оценка стоимости операций по достижению цели

Цель 3 — получение предварительной информации о переменных технологического процесса.

Для достижения этой цели используются малые выборки «настроечных» опытов, которые проводились при реализации цели 1, либо были специально поставлены (обычно информацию получают пассивно, наблюдением). Задача решается методами математической статистики и теории случайных функций. Находятся математические ожидания, дисперсии (моменты) переменных, их интервальные оценки; оценивается закон распределения и рассчитываются коэффициенты корреляции переменных; оцениваются стационарность, эргодичность случайных функций, рассчитываются корреляционные функции и др. Схема реализации этой цели изображена на рис. 1.12, в. Системные элементы достижения цели приведены в табл. 1.5.

Таблица 1.5.

Системные элементы достижения цели 3

Элементы ЯИ	Содержание элементов
A_3, C_3	Задан способ производства, сырье, катализатор; имеется лабораторная установка, на которой получены малые выборки пассивной информации
M_3	Модели математической статистики и теории случайных функций и постановка пассивного плана эксперимента
T_3	Принятие решений по статистическим критериям
K_3	Количественные нормы статистических критериев (t -критерий, критерий Пирсона и др.)
R_3	Все статистические оценки переменных, оценка динамики объекта и связей между переменными

Цель 4 — отсеивание факторов.

В числе факторов, предложенных для отсеивания, могут быть и дискретные, и непрерывные. Цель достигается постановкой экспериментов на лабораторной установке.

Задача решается методами экспертных оценок, планов дробного факторного эксперимента и Плакетта — Бермана, методом случайного баланса. Планы строятся по критерию минимального числа опытов. Схема реализации цели 4 изображена на рис. 1.12, а системные элементы — в табл. 1.6.

Таблица 1.6.

Системные элементы достижения цели 4

Элементы ЯИ	Содержание элементов
A_4, C_4	Лабораторная установка, список факторов, их статистические оценки
M_4	Модели разделения факторов на значимые, относящиеся к «шумовому» фону, модели линейной регрессии и планы эксперимента по их получению
T_4	Принятие решений на основе статистических критериев
K_4	Количественные нормы статистических критериев (обычно t -критерий)
R_4	Факторы, существенно влияющие на выходную переменную; области существования факторов, интервалы варьирования, оценочное число параллельных опытов

Цель 5 — интерполяция, экстраполяция и оптимизация по математическим моделям полиномиального вида первого, второго и более высоких порядков. Рассмотрим несколько подцелей этой цели (схема реализации цели 5 изображена на рис. 1.12, *д*, а системные элементы в табл. 1.7).

Таблица 1.7.

Системные элементы достижения цели 5

Элементы ЯИ	Содержание элементов
A_5, C_5	Лабораторная установка; список переменных, существенно влияющих на ее работу, их характеристики
M_5	Математические модели полиномиального вида первого, второго и более высоких порядков и планы эксперимента по их получению
T_5	Принятие решений на основе статистических критериев
K_5	Количественные нормы статистических критериев (обычно <i>t</i> - и <i>F</i> -критерии, критерий Кохрена)
R_5	Интерполяционное и экстраполяционное значения выходной переменной, экспериментальное значение, частный и локальный оптимум функции цели; минимальные ресурсы объекта исследования

Цель 5а — интерполяция результатов исследования. Если область эксперимента достаточно мала, то можно использовать линейную модель, в противном случае используется нелинейная модель. Задача математически формулируется так: имеется вектор факторов X и выходная переменная объекта y ; область эксперимента — $A \leq X \leq B$; необходимо найти интерполяционное значение y_n по $X_n \in X$. Задача решается постановкой ряда опытов по планам ПФЭ, ДФЭ, ЦКОП, ЦКРП, Шеффе, обработкой результатов..

Цель 5б — экстраполяция результатов исследования. Условия те же, что и в предыдущем опыте, но ищется экстраполяционное значение y_n по $X_n \in X$, т. е. по значениям факторов, лежащим вне области эксперимента. Задача решается упомянутыми выше методами (см. цель 5а). Здесь часто используется каноническое преобразование модели и требуется обязательная экспериментальная проверка y_n^* .

Цель 5в — получение экстремума функции цели в области эксперимента. Для достижения этой цели могут использоваться как линейная, так и нелинейная модели, которые получаются уже упоминавшимися при достижении цели 5а планами эксперимента. Задача решается методом линейного программирования (по линейной

модели) или любым нелинейным методом (симплексным, методом Гаусса, ридж-анализом и др.). Решение $y_{\text{опт}}$ находится либо в области эксперимента $X_{\text{опт}} \in X_3$, либо на границе области.

Цель 5г — *получение частного и локального оптимумов объекта исследования.* Для достижения этой цели чаще всего используется линейная модель, полученная по планам ПФЭ (ДФЭ). Частный (промежуточный) оптимум и локальный (один из возможных оптимумов объекта исследования по области существования факторов) обычно находят одним из поисковых методов оптимизации, например, методом крутого восхождения. Поиск осуществляется за пределами области эксперимента. Полученное $y_{\text{опт}}$ обязательно уточняется экспериментально.

Цель 5д — *минимизация ресурсов объекта исследования.* Эта цель достигается решением задачи, обратной цели 5б: находят X_{min} при заданном y . Модель может быть как линейной, так и нелинейной. При решении задачи возможны различные варианты: стабильны все факторы X_i кроме одного, который и нужно минимизировать; стабильны все факторы X_i , кроме двух, один из которых должен быть минимизирован, а для другого — допускается полный расход ресурса ($X_j = -1$); переменны все факторы X_i , сумму которых необходимо минимизировать. Часто задача решается графоаналитическими методами.

Цель 6 — *получение математических моделей объекта исследования.* Достижение этой цели осуществляется неформальным (и, по-видимому, неформализуемым) предложением вида (класса) математической модели исследуемого процесса исходя из имеющейся у исследователя информации, его творческого мышления и интуиции. Чаще всего модель предлагается в виде

$$\dot{Y} = \frac{dY}{dt} = f(Y(t), U(t), A(t), \varepsilon(t), t),$$

где $A(t)$ — l -мерный вектор неизвестных параметров системы, состоящей из коэффициентов системы дифференциальных уравнений; $\varepsilon(t)$ — помехи объекта и измерительных приборов; t — время.

Необходимо по результатам наблюдения (или по плану) за $y(t)$, $u(t)$ найти вектор параметров $A(t)$. Задача решается методами идентификации. В результате получают численные оценки параметров $A(t)$, т. е. собственно математическую модель объекта в заданном классе моделей (см. рис. 1.12, е и табл. 1.8).

Таблица 1.8.

Системные элементы достижения цели 6

Элементы ЯИ	Содержание элементов
A_6, C_6	Оптимальные переменные объекта исследования, вид математической модели процессов объекта, лабораторная установка
M_6	Математические модели идентификации
T_6	Принятие решений об адекватности моделей, учет априорной информации и экспериментальная проверка
R_6	Количественные оценки параметров математической модели (математическая модель объекта)

Цель 7 — отсеивание (дискриминация) математических моделей. Достижение этой цели возможно двумя комплексами операций:

1) постановкой эксперимента в некоторой области $A \leq X_3 \leq B$; расчетом параметров для конкурирующих моделей по этим опытам; прогнозом значений y по всем моделям для некоторого шага факторов в области эксперимента; расчетом критерия (Рота, Бокса и Хилла и др.), характеризующего различие в моделях; постановкой следующего опыта в той точке факторного пространства, где различие между моделями было наибольшим; включение этой точки в массив данных и повторение всех операций до тех пор, пока значение критерия не будет изменяться;

2) расчетом критериев математической модели (сумм квадратов отклонений, функций максимума правдоподобия, байесовой оценки); анализом этих критериев и принятием решений по выбору наилучшей в смысле применявшегося критерия модели. Таким образом, задача дискриминации модели решается последовательным поиском наилучшей точки экспериментальной области, где различие между моделями наибольшее, и расчетом в этой точке некоторых критериев моделей, по которым судят об их пригодности (см. рис. 1.12, ж, табл. 1.9).

В результате получают одну математическую модель (модели) объекта исследования, которую и используют при проектировании непрерывных технологических процессов (комплексов).

Таблица 1.9.

Системные элементы достижения цели 7

Элементы ЯИ	Содержание элементов
A_7, C_7	Несколько математических моделей для одного объекта исследования и возможность постановки экспериментов на лабораторной установке
M_7	Модели определения максимального различия между математическими моделями и модели их дискриминации
T_7	Принятие решений по специальным критериям
K_7	Количественные оценки специальных критериев дискриминации
R_7	Одна, наилучшая в смысле применявшихся критериев, математическая модель объекта исследования

1.7. Введение в методы исследования

Методика исследования — это совокупность способов, приемов исследования. Она отвечает на вопрос: как, какими способами проводить исследование.

Методика, которая относится ко всему исследованию и представляет стержневые способы и приемы, проходящие через все исследование, является общей. Для отдельных опытов (в пределах исследования) или серии опытов могут потребоваться специфические способы исследования. В этих случаях необходимо разработать частные методики, отличные от общей.

Значение частных методик возрастает с увеличением разнообразия явлений или сторон явления, подлежащих изучению.

Методика исследования определяет оборудование, процессы исследования, этапы исследования, количество опытов, план работы (планирование эксперимента), затраты времени и средств.

Правильная методика — залог успеха исследования. Неметодичность исследования или неверная методика приводят к ошибкам, в результате чего накапливаются груды «сырого» материала, из которого никаких выводов сделать нельзя.

От наблюдения к обобщению и от обобщения к проверке в жизни, на практике — таков общий путь исследования. Эти три ступени обязательны для законченного исследования.

Первая ступень — наблюдение, взятое в широком смысле, охватывает пассивную его форму (наблюдение без вмешательства в условия развития явления) и активную (наблюдение исследователем за развитием явления в определенных условиях, эксперимент). Но,

прежде чем более подробно ознакомиться с проведением наблюдений, следует изучить широко применяемые средства наблюдения — **измерения и отметки**.

1.7.1. Измерения и отметки

Задача научного исследования состоит в том, чтобы установить закономерности развития явлений. При определении закономерности мы ищем постоянное и необходимое отношение, связь между явлениями, лежащую в основе их развития. Для подавляющего большинства исследований закономерности представлены численными величинами, полученными в результате измерений.

Здесь и далее **величина** — **обобщение конкретных понятий** (длины, массы, силы, скорости и т.п.) **свойств физических тел**.

Измерение есть процесс сравнения данной величины с ее значением, принятым за единицу. Всегда надо стремиться измерять величины в процессе развития явления, а не ограничиваться их измерениями «до» и «после», так как познавательное значение последних сравнительно мало. Общая классификация измерений приведена в таблице 1.10.

Таблица 1.10

Классификация измерений

Класс, признак	Подклассы, виды
1. Количество однотипных измерений	1-1. Разовые (однократные) 1-2. Многоразовые (неоднократные) 1-2-1. Равноточные 1-2-2. Неравноточные 1-2-3. Дискретные 1-2-4. Непрерывные
2. Отношение к измеряемой величине	2-1. Прямые (непосредственные) 2-2. Косвенные (посредственные)
3. Связь измерений	3-1. Синхронные (по времени или пути) 3-2. Асинхронные

Особенности измерений различных видов характеризуются далее при рассмотрении последних.

Чем глубже исследование, чем большие проблемы оно решает, тем совершеннее и точнее должны быть способы измерений и измерительная аппаратура (табл. 1.11).

Таблица 1.11

Измеряемые величины и основные инструменты и приборы для измерений

Определяемые величины	Аппаратура
<p>Геометрические: длина, высота, глубина, ширина, углы, профили, площади, объемы</p>	<p>Мерительные ленты, линейки, шупы, штангенинструменты; светодальномеры; микрометры, нутромеры, глубиномеры; миниметры, оптиметры, интерферометры, индикаторы, проекторы, микроскопы, пневматические и электрические измерительные приборы; планиметры, профилометры, эклиметры; тензомеры; счетчики объемов, уровнемеры, расходомеры; специальные датчики и регистрирующие устройства</p>
<p>Кинематические: скорость, ускорение, колебания, повторяемое движение</p>	<p>Измерители пути (угла поворота) и времени, измерители линейной и угловой скоростей, стробоскопы, тахометры, тахогенераторы, счетчики оборотов; ускорениемеры; торсиографы, вибрографы, частотомеры и амплитудомеры; специальные датчики и регистрирующие устройства</p>
<p>Динамические: масса, импульс, сила, работа, момент, мощность</p>	<p>Весы, импульсомеры, динамометры, тормозные и другие нагрузочные устройства, манометры, барометры, индикаторы, работомеры; специальные датчики и регистрирующие устройства</p>
<p>Некоторые другие величины: время, световые величины, цвет, температура; электрические, химические и структурные свойства твердых, жидких и газообразных тел; излучение; звуковые величины</p>	<p>Часы, секундомеры, вибрационные и электронные устройства; фотометры, нефелометры, дымомеры, люксметры, калориметры, хроматографы, спектрометры; термометры, термоэлементы, пирометры, омметры, вольтметры, амперметры, ваттметры, гальванометры; анализаторы, полярографы; почвенные плотномеры, твердомеры (Бринелля, Роквелла), склероскопы Шора, вискозиметры, пенетрометры; оптические микроскопы, электронный микроскоп, ультразвуковые приборы, радиометрическая аппаратура (счетчики, камеры, приемники с пересчетными устройствами и автоматической записью); шумомеры, дециомеры, звукоанализаторы; приемники инфракрасных излучений</p>

Большую часть перечисленной аппаратуры выпускает промышленность, но во многих случаях для специфических измерений исследователь должен сам конструировать приборы, используя в готовом виде лишь отдельные узлы заводских приборов. **Отметим основные тенденции в развитии способов измерений:** повышение

точности измерений; переход к измерениям микровеличин; применение приборов безынерционных и без трения или по крайней мере с величинами сил инерции и трения, которыми практически можно пренебречь; применение бесконтактных приборов; регистрация непрерывно изменяющихся величин; уменьшение затрат времени и труда на производство измерений; синхронные измерения; получение наибольшего количества надежной информации.

Изложенные здесь краткие указания о выборе способов измерений рассчитаны лишь на самые общие сведения и не могут заменить описаний конструкций приборов, их применения и условий выбора той или иной аппаратуры, помещенных в специальной литературе или в заводских руководствах.

Как правило, для измерений какой-либо неизменной (или близкой к постоянной) величины, например диаметра вала по линии сопряжения осевой и нормальной плоскостей или массы детали, применяют наиболее простые способы. В этих случаях измерительный инструмент или прибор, например концевые меры, микрометр, линейку, мерную ленту, выбирают в зависимости от измеряемой величины, удобства пользования и необходимой точности измерений.

При неоднократных измерениях изменяющихся величин лучше употреблять считающие или записывающие приборы и устройства.

Считающие приборы показывают интегральный, суммарный результат измерений какой-либо одной величины, например суммарное количество оборотов вала за время измерения. В тех случаях, когда результаты измерений одной изменяющейся величины желательно статистически обработать, рекомендуется использовать статистические считающие приборы или классификаторы. В этих приборах значения переменной величины заранее разбивают на классы. Как только она достигает значения, соответствующего заданному классу, датчик, т. е. чувствительный элемент прибора, посылает импульс в счетчик этого класса, и он суммирует количество этих импульсов. В результате на ряде счетчиков прибора будет показано статистическое распределение измерений различных значений переменной величины. Например, в обычном тяговом циферблатном динамометре стрелку прибора заменим движущимся контактным устройством, а шкалу показаний разделим на классы и по последним поставим неподвижные контакты, соединив каждый из них со счетчиком электрических импульсов. Тогда при включении в эту систему источника питания и при воздействии на силовое звено динамометра переменной нагрузки каждый из счетчиков будет

показывать количество замыканий, соответствующих определенному диапазону нагрузок.

Статистические считающие приборы сравнительно просты, надежны, установить их несложно. Их основное достоинство — возможность регистрации данных в течение более длительного времени по сравнению с обычными записывающими приборами. Недостаток их состоит в том, что они не регистрируют мгновенные функциональные зависимости и связи между несколькими переменными.

В случае функционального анализа связей между несколькими переменными требуется синхронная запись их значений. Например, надо установить зависимость нагрузки на вал кривошипа от угла поворота вала. Очевидно, для этого необходимо знать значения нагрузки и угла поворота вала в каждое данное мгновение, то есть синхронно записывать одновременно изменяющиеся переменные величины.

Недостаток такой регистрации измерений — сложность приборов и большая трудоемкость обработки записей. Кроме того, узлы, входящие в синхронизирующую систему (от датчика до записывающего устройства), должны изготавливаться с высокой точностью, так как в противном случае записи искажаются. Например, при сдвиге фаз гармонических колебаний ошибка может быть очень большой.

Существуют три метода синхронной записи нескольких переменных величин: при помощи осциллографа, запись на магнитный носитель, видеосъемка. При сложных явлениях протекание различных переменных регистрируют синхронно тремя или двумя методами.

Наиболее распространена синхронная запись при помощи различного типа осциллографов. При необходимости осциллографы блокируют, одновременно включая, выключая и подавая сигналы времени, начала и конца опыта из одного сигнального устройства.

В зависимости от характера измеряемой величины датчики для осциллографов могут различаться по конструкции и принципам действия. Главные требования, предъявляемые к датчикам для осциллографической записи: надежность; устойчивость показаний по времени (при измерении одной и той же величины) и при обычных для опытов колебаниях температуры; удобство монтажа; надежность соединения с проводами осциллографа или усилителя.

Для определения величин или связанных с деформацией детали, или могущих деформировать датчик, достаточно простыми, удобными и точными являются проволочные и фольговые датчики. Проволочные

датчики делают из специальной проволоки (константановой и др.). Датчики из тонкой проволоки компактны, но требуют усиления посылаемых ими сигналов; датчики из более толстой проволоки или фольговые (так называемые мощные) при надлежащей чувствительности шлейфов можно применять без усиления, но они менее компактны и монтировать их труднее.

Основой проволоочного датчика служит зигзагообразно уложенная проволока (рис. 1.13, *а*), называемая иногда решеткой; при деформации проволоки изменяется ее удельное сопротивление, что и отмечает осциллограф.

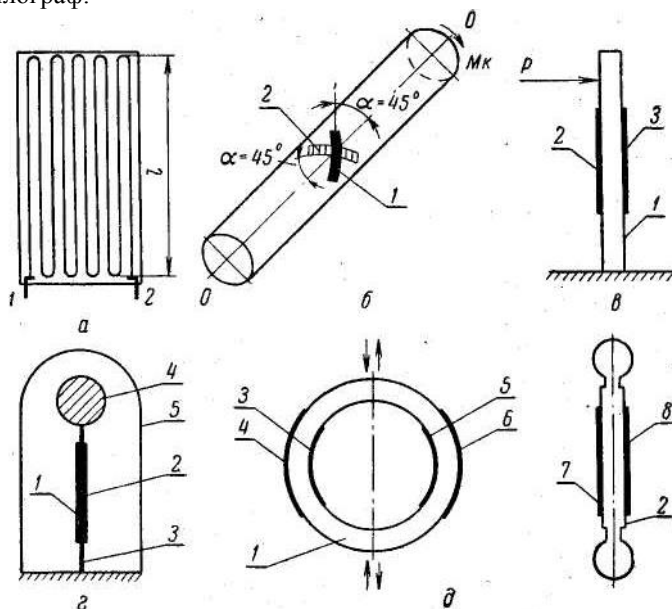


Рис. 1.13. Схемы проволоочных датчиков и их использование:

а —общая схема датчика: *l*— база датчика; 1 и 2— выводы; *б* — расположение проволоочных датчиков для измерения крутящего момента M_k : 1— датчик, работающий на сжатие; 2— датчик, работающий на растяжение; *в* — измерение упругих деформаций: 1 — деталь, подвергающаяся упругой деформации под воздействием силы P ; 2 — датчик, работающий на растяжение; 3 — датчик, работающий на сжатие (такого же типа датчик может отмечать микроперемещения); *г*—устройство для измерения ускорений: 1 и 2—датчики; 3 — элемент, на который наклеены датчики; 4 — масса; 5 — стакан с демпфирующим составом; *д* —динамометрические: 1 — кольцо; 2 — палец; 3—8 — датчики, работающие на растяжение или на сжатие, в зависимости от действия сил, а в кольце — в зависимости от места наклейки.

Такую проволочную решетку наклеивают продольной осью вдоль оси напряжения на исследуемую или на специально сделанную деталь, деформация которой достаточна для действия датчика. При наклейке принимают все меры к тому, чтобы правильно расположить датчик, электрически изолировать его от детали, предохранить от влаги и механических повреждений, и также к температурной компенсации. Концы датчика включают в измерительную цепь обычно по мостовой схеме. На рисунке 1.13, б, в, г и д схематически показаны примеры использования проволочных датчиков в некоторых, наиболее часто встречающихся случаях. При измерении проволочными датчиками различных переменных величин вращающихся деталей необходимо применять специальные токосъемные устройства.

Если на вал (рис. 1.13, б) действует не только крутящий, но и изгибающий момент, оппозитно двум показанным датчикам наклеивают еще два так, чтобы в каждой паре датчиков, находящихся в одной ветви моста, возникали деформации одного знака и равновесие моста не нарушалось. Такая схема одновременно компенсирует и колебания температуры. Полупроводниковые датчики на 1—2 порядка чувствительнее проволочных. При измерении достаточно больших угловых и линейных перемещений небольшой частоты применяют контактные и реостатные (реохордные) датчики (рис. 1.14, а, б, в); последние могут быть использованы также для измерений ускорений.

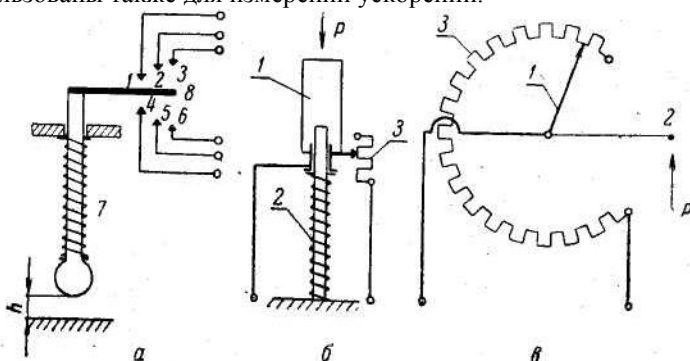


Рис. 1.14. Контактные и реостатные датчики:

а — контактный датчик: 1—6 — контакты датчика; 7 — возвращающая пружина; 8 — контакт; *h* — измеряемый размер; *б* — реостатный датчик: 1 — деталь, воспринимающая усилие P ; 2 — возвращающая пружина; 3 — реостат (может быть реохорд); *в* — реостатный или реохордный датчик: 1 — вращающийся контакт; 2 — рычаг, приводящий в движение вращающийся контакт; 3 — реостат (реохорд).

Фотоэлектрические датчики для измерения крутящих моментов не требуют усилителей и токосъемных устройств, но сложны по конструкции.

В некоторых случаях для измерений давлений, перемещений (в пределах нескольких миллиметров), скоростей, вибраций и крутящего момента при частоте колебаний измеряемой величины в сотнях герц можно применить индукционные датчики, а некоторые из них использовать для записи сигналов осциллографом без усилителя.

Иногда одним шлейфом осциллографа можно одновременно записывать две величины, если они изменяются в определенных соотношениях. Например, если сила связана с перемещением, то кривая может быть одна, а масштабы для определения силы и перемещения — разные. Этим экономят шлейфы.

При высоких частотах колебаний измеряемой величины, например при записи процессов в камерах сгорания поршневых двигателей, целесообразно применять пьезоэлектрические или емкостные датчики и электродные (катодные) осциллографы. Обычно ими измеряют одновременно только 2—3 величины.

Синхронизировать измерения при помощи двух электронных осциллографов можно одновременной киносъемкой их экранов на пленку достаточно высокой чувствительности.

Недостаток киносъемочной синхронизации — большие затраты времени на расшифровку записей, так как в большинстве случаев приходится показания приборов в каждом кадре переписывать в таблицы, а иногда строить по этим записям кривые.

Современные скоростные киноаппараты (от 150 до 4000 снимков в 1 сек) приемлемы для исследований; их можно использовать не только для синхронизации записей на экранах катодных осциллографов, но и для регистрации быстротекущих видимых явлений синхронно с записями магнитоэлектрическим осциллографом. Таким образом достигается регистрация на киноленте внешних признаков синхронно с регистрацией сил, напряжений или других факторов и функций, что значительно увеличивает количество получаемой в опыте информации и этим повышает надежность исследования.

Синхронизация одновременной записью ряда переменных величин на магнитную ленту при помощи частотного модулятора и магнитофона удобна простотой аппаратуры и тем, что на одну ленту одновременно записывают сигналы нескольких датчиков. Есть приборы, которые дают сразу среднее значение колеблющейся

измеряемой величины. Несмотря на то, что показания этих приборов не требуют дополнительной обработки, следует осторожно относиться к их применению, так как многие процессы при исследовании объектов характеризуются не только средними величинами, но и отклонениями от средних.

Существуют устройства, трансформирующие сигналы датчиков и записывающие производные или интегралы изменяющихся величин. При выборе таких устройств учитывают их надежность, вносимые ими ошибки и сложность их эксплуатации.

Число однотипных приборов, применяемых в исследовании, зависит от целей последнего. Например, при динамометрировании (линейном, в плоскости и пространственном) можно измерять соответственно только силу, две и три ее составляющих. Полную информацию о положении твердого тела в пространстве дают три координаты линейных перемещений и три координаты углов поворота. Помимо измерений, в экспериментах делают и отметки, цель которых состоит в количественной и качественной фиксации явления или его сторон во всей их сложности. Простейшей отметкой является запись в виде замечаний экспериментатора о всех интересных фактах в развитии явления. Этими записями никогда не следует пренебрегать, так как они помогают «осветить» стороны явления, которые вначале не считались первостепенными, а потом стали важными, а также прояснить некоторые их особенности. Эти же записи содержат новые мысли, соображения и замечания. Бывают и другие отметки: фотография, звукозапись, фиксация состояния объекта нафталиновым сплавом, а лезвий и поверхностей - пластмассами, гипсом, легкоплавким металлом, органической пленкой; эскизирование, сбор коллекций и отбор качественных проб; статистика явлений и т. д.

1.7.2. Точность измерений

Ошибки измерений. Чем выше точность измерений, тем надежнее результаты исследования. Оценка точности и надежности обязательна, так как полученные значения могут лежать в пределах возможной ошибки эксперимента, а выведенные закономерности — оказаться неясными и даже неверными.

Точность есть степень соответствия результата измерений действительному значению величины. Понятие точности связано с понятием ошибки: чем выше точность, тем меньше ошибка, и наоборот. Самые точные приборы не могут показать действительного значения величины, их показания содержат ошибку.

Абсолютная ошибка — это разность между действительным значением измеряемой величины x и ее наблюдаемым значением a .

Практически под абсолютной ошибкой Δ понимают разность между результатом измерения при помощи более точных методов или приборов высшей точности (образцовых) и значением этой же величины, полученным прибором, применяемым в исследовании:

$$\Delta = x - a \approx a_{\text{обр}} - a. \quad (1.26)$$

Практически понятие «точность» имеет несколько определений. Иногда под точностью понимают значение абсолютной ошибки. Можно сказать, например, что точность измерения данным инструментом равна 0,01мм. Чаще под точностью понимают безразмерную величину или отношение абсолютной ошибки к действительному значению величины (это отношение иначе называется относительной ошибкой)

$$\Delta_1 = \pm \frac{\Delta}{x}, \quad (1.27)$$

или выражают ее в процентах к действительному значению величины

$$\Delta_2 = \pm \frac{\Delta}{x} 100\% \quad (1.28)$$

Для современных способов измерений можно считать, что величина Δ незначительна в сравнении как с действительным значением x величины, так и с наблюдаемым при помощи наиболее точных (образцовых) приборов ее значений $a_{\text{обр}}$. Допуская, что при измерении образцовыми приборами x приблизительно равна $a_{\text{обр}}$, и считаясь с тем, что мы не можем измерить действительное значение x , выразим точность так:

$$\Delta_1 = \pm \frac{\Delta}{a_{\text{обр}}}, \quad (1.29)$$

$$\Delta_2 = \pm \frac{\Delta}{a_{\text{обр}}} 100\% \quad (1.30)$$

Так, например, если $\Delta = 0,01$ мм и $a_{\text{обр}} = 10$ мм, то точность измерения

$$\Delta_1 = \pm \frac{0,01}{10} = \pm 0,001;$$

$$\Delta_2 = \pm \frac{0,01}{10} 100 = \pm 0,1\%.$$

Ошибки могут возникать из-за несовершенства измерительного прибора, влияния условий его эксплуатации и опытности наблюдателя. Неточная градуировка, трение и зазоры в механизме, инерционность устройства, износ, старение или усталостные явления и другие аналогичные факторы определяют ошибку прибора. Условия эксплуатации; расположение прибора, его наладка, температура среды и ее колебания; электрические магнитные поля и их изменения, давление воздуха, влажность, напряжение, частоты и другие внешние факторы, большей частью переменные, определяют ошибки измерений. Величина ошибки результата, зависящая от наблюдателя, связана с его опытностью, внимательностью, быстротой реакции, особенностями зрения и другими личными качествами.

В зависимости от того, известны ли причины разных ошибок или нет, можно ли их определить и оценить их размеры и направление, все ошибки разделяют на два класса: *систематические и случайные*.

Систематическими называют такие ошибки, которые возникают из-за известных причин, действующих по определенным законам и, как правило, в определенном направлении. Таким образом, систематическими являются те ошибки или погрешности, которые можно изучить и определить количественно. Их необходимо знать и вносить соответствующие поправки в показания приборов, то есть эти ошибки можно рассматривать как поправки к показаниям приборов. Обычно поправки, связанные с конструкцией приборов, указывают в их паспортах.

Систематические ошибки может определить и сам исследователь, что особенно необходимо для приборов, им конструируемых. Находят систематические ошибки, калибруя измерительные приборы. Систематические ошибки иногда можно обнаружить, сопоставляя опытные данные с изменяющимися внешними условиями; соответствие между изменениями опытных данных и этих условий указывает на наличие систематической ошибки. При измерениях необходимо предупреждать возникновение непредусмотренных ошибок. Для этого надо тщательно устанавливать измерительную аппаратуру, например по отвесу или уровню, проверять, не смещен ли указатель с нулевой точки, не возникают ли дополнительные самоиндукция и емкость при том или ином расположении проводов или кабеля. Если систематические ошибки могут быть вызваны внешними условиями (переменной температурой, тряской, осадками), следует компенсировать их влияние или измерять только при более или менее устойчивых условиях.

Проверка и калибровка приборов. Исправность приборов проверяют периодически; для того чтобы выявить их ошибки и, следовательно, повысить точность измерений, их калибруют. При калибровке одну и ту же величину измеряют сначала проверяемым прибором, а затем более точным и устанавливают разность Δ между этими измерениями. Таким образом, при калибровке показания приборов, выбранных для измерений, сверяют с показаниями приборов более высокой точности и устанавливают погрешности прибора. Этой же цели служат калибровочные кривые (кривые показаний проверяемого прибора в функции показаний прибора высокой точности). Чтобы точнее определить ошибки, следует сравнивать показания проверяемого и образцового приборов при разных знаках изменения измеряемой величины и в достаточно широком диапазоне нагрузок. Например, если реостатный или реохордный датчики с силовым звеном в виде пружины проверяют эталонными гирями, целесообразно в определенном порядке нагружать силовое звено этими гирями и затем разгружать его, снимая сначала гирю, поставленную последней, затем предпоследней и т. д. Из-за трения, зазоров в сочленениях механизма и других факторов данные калибровки при увеличении нагрузки и при разгрузке могут расходиться (рис. 1.15,а).

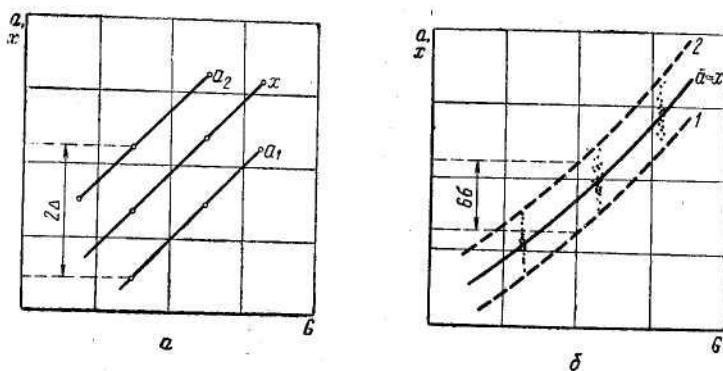


Рис. 1.15. Результаты калибровки прибора:

a — по образцовому прибору; x — показания образцового прибора; a_1 — показания проверяемого прибора при нагрузке и a_2 — при разгрузке; b — с разбросом результатов из-за воздействия неизвестных факторов; $\bar{a} \approx x$ — средняя ряда измерений; 1, 2 — нижняя и верхняя границы результатов; σ — величина разброса результатов калибровки,

В этом случае кривые калибровки эквидистантны и расстояние между ними (по ординате) равно 2Δ , в то время как показания образцового прибора, отмечающего величину, близкую к истине, расположены на линии x , находящейся между линиями a_1 и a_2 . Отсюда, если калибруемый прибор используется только в одном из режимов — нагрузки или разгрузки, систематическая ошибка равна Δ (по рис. 1.15, а) с соответствующим знаком, и ее можно учесть в результате измерения. Если кривые не эквидистантны, ими пользуются для внесения различных по величине поправок при обработке опытных данных. В том же случае, когда калибруемый прибор работает в двух режимах (нагрузки и разгрузки), знак ошибки в большинстве случаев учесть трудно или невозможно, и систематическая ошибка превращается в случайную величину $\pm\Delta$.

Бывают также случаи, когда калибровку проводят методами, не устраняющими влияния неизвестных случайных причин, так что неизбежен разброс точек результатов каждой ступени нагрузки (см. рис. 1.15, б). В таком случае проводят сглаженную среднюю $\bar{a} \approx x$. Кроме того, по крайним точкам отклонений от средней \bar{a} также проводят сглаженные кривые 1 и 2 (в общем случае неэквидистантные), расстояние между которыми по ординатам характеризует удвоенную максимальную статистическую ошибку прибора при каждой данной ступени нагрузки. Таким образом, здесь определяется наибольшая случайная ошибка прибора $\pm 3\sigma$. Систематические ошибки можно выявить и другими методами:

а) проводить измерения два раза так, чтобы причина погрешности первого измерения оказала противоположное действие на результат второго измерения;

б) строить график результатов измерений постоянной величины в функции аргумента (например, времени) в случае прогрессирующей погрешности (например, при падении напряжения);

в) заменять измеряемый объект известной мерой, находящейся в равных с ним условиях (сравнение измеряемого параметра с известной мерой).

Перед проверкой аппаратуры следует выяснить, не будут ли на нее действовать какие-либо факторы, дающие дополнительную систематическую ошибку. Если это так, прибор проверяют после установки его для измерений. Например, если у датчика, смонтированного на машине, нулевая точка отклонится, то калибровать его надо после установки.

Приборы следует проверять перед измерениями, иногда в течение опытов и обязательно по окончании их, чтобы можно было

судить о точности измерений. В связи с этим работу с приборами и инструментом надо организовать так, чтобы можно было регулярно проверять их показания. Непроверенными приборами пользоваться нельзя.

Если исследователь сам сконструировал прибор, помимо калибровки и градуировки, надо установить его эксплуатационные возможности и характеристики: чувствительность, диапазон измерений, влияние неизмеряемых факторов и т. д.

Случайными ошибками называются те, причины которых неизвестны и которые невозможно учесть заранее. Такие ошибки характеризуют точность измерения. Суммарную случайную ошибку вычисляют на основе теории ошибок.

При достаточной чувствительности прибора и правильности приемов измерения случайные ошибки проявляются в том, что при ряде измерений одной и той же величины получаются разные результаты отсчетов в последних значащих цифрах. Хотя абсолютная величина и знак ошибок случайны, эта величина не может превысить некоторой предельной, зависящей от условий измерения. Полностью исключить случайные ошибки нельзя, но исследователь обязан определить возможную ошибку опыта, вычислений или исследования, чтобы установить их точность.

Случайные ошибки измерений можно выразить несколькими понятиями. Часто применяют понятие *предельной* ошибки, под которым подразумевают наибольшую случайную ошибку при правильном пользовании исправным прибором и устранении систематических ошибок при внесении соответствующих поправок, то есть такое число Δ_n , при котором $x - a \leq \Delta_n$.

Если установленную систематическую ошибку по каким-либо причинам невозможно применить в качестве поправки к результату измерений, предельная ошибка увеличивается на величину этой систематической ошибки.

Таким образом, значение измеряемой величины можно представить формулой:

$$x = a \pm \Delta_n, \quad (1.31)$$

что следует читать так: истинное значение измеряемой величины находится в пределах от $a - \Delta_n$ до $a + \Delta_n$.

Зная предельную абсолютную ошибку Δ_n можно вычислить также предельную относительную ошибку соответственно формулам (1.29) и (1.30):

$$\Delta_1 = \frac{\Delta_n(a)}{a} \quad (1.32)$$

или

$$\Delta_2 = \frac{\Delta_{\Pi}(a)}{a} 100\%. \quad (1.33)$$

Чем больше значение измеряемой величины, тем меньше относительная ошибка, поэтому относительную предельную ошибку справочных данных обычно относят к наибольшему значению величины, которую можно получить при измерениях данным прибором.

При неравномерных шкалах следует учитывать, что предельная ошибка будет различной в разных местах шкалы.

Предельную ошибку прибора устанавливают, изучая и проверяя его до измерений. Эта ошибка может после опытов измениться только в двух случаях: если после измерений при проверке обнаружено изменение погрешности прибора (в данном случае измерение повторяют или принимают наиболее вероятный закон изменения погрешности за время измерений и вносят в последние соответствующие поправки) и если при неравномерной шкале перед измерениями предполагали отсчитывать показания в одной части шкалы, а фактически результат измерения получали в другой части шкалы, с иной точностью. В таблице 1.12 приведены предельные ошибки некоторых приборов.

Из таблицы видно, что при пользовании осциллографом с усилителем ошибка измерений может возрасти. Ошибки могут быть большими при осциллографической записи измерений нескольких величин в движущейся по полю машине. Эти ошибки во многом зависят от наведения (индукции) токов. Основные меры борьбы с такого рода искажениями измерений: применение мощных датчиков, не требующих усиления сигналов при записи; экранирование проводов; уменьшение длины проводов; поддержание неизменного расстояния между проводами, а также между проводами и грунтом. Следует предпринять также все необходимые меры против искажений, вызываемых механическими сотрясениями деталей осциллографов.

Таблица 1.12

Предельные ошибки при различных способах измерений

Способ измерения, инструмент, прибор	Предельная ошибка, % к верхнему (наибольшему) пределу измерения
Стальная 20-метровая лента	0,20— 0,30
Планиметры линейные и полярные	0,40— 0,70
Интерферометры	1,00—10,00
Угломеры оптические	0,5 — 2,00
Светодальномеры	0,01— 0,02
Тахометры центробежные	0,40— 2,50
Тахогенераторы	2,50— 4,00
Тахоскопы	0,80— 1,10
Стробоскопы	0,10— 1,00
Весы:	
торговые и автомобильные	0,40— 1,20
технические	0,10— 0,20
аналитические	0,0001— 0,01
Динамометры тяговые:	
пружинные	1,00— 3,50
гидравлические	0,70— 2,00
электрические (датчики без усилителя)	0,20— 0,50
Тормоза:	
балансирные	0,40— 1,00
гидравлические	1,50— 2,50
пневматические	1,50— 4,00
Манометры с трубкой Бурдона	1,00—10,00
Ртутные манометры	1,00— 2,50
Фотоэлектрические датчики	0,40— 2,00
Стандартные секундомеры	0,40— 0,70
Термопары (без усилителя)	0,50— 2,50
Фотометры	0,05— 2,00
Спектрометры	0,10— 5,00
Полярграфы	1,00— 5,00
Газоанализаторы с поглощением	0,50— 5,00
Хроматографические газоанализаторы	0,10— 2,00
Дифференциальные калориметры	0,01— 1,00
Стандартные вискозиметры	1,00— 4,00
Ртутные технические термометры	0,30— 2,00
Полупроводниковые термометры	0,10— 1,00
Твердомеры ударного действия	7,00—15,00
Запись осциллографом при усилении	1,50— 4,50
Магнитная запись при частотной модуляции	2,00— 5,00
Пневматические и масляные калибраторы	1,00—10,00

Рассмотрим, как применить изложенное при подборе аппаратуры в наших схематических примерах.

В первом примере основным фактором, определяющим показатели работы плуга, служит степень затупления лемехов. Нет смысла синхронно записывать степень затупления и соответственные изменения показателей плуга, так как затупление лемехов — длительный процесс. Подбираем комплекты лемехов одной и той же степени затупления — от новых до наиболее изношенных. Затупление лемеха можно измерить двумя способами: по ширине задней фаски и по величине секущей, отделяющей затупленную часть лезвия. Принимая для контроля оба способа, затупление лемехов измеряем прецизионной линейкой и зубомером. Сопrotивление плуга определяем динамографом, глубину вспашки — линейкой, захват плуга — метром, время — секундомером, пройденный путь — 20-метровой стальной лентой.

Иначе следует подбирать аппаратуру во втором примере. Здесь синхронно надо измерить пройденный путь, время, расстояние от оси подшипника колеса (по крайней мере, одного) до основания, по которому колесо перекачивается, прогиб ножевого бруса в его середине и расстояние от середины ножевого бруса до почвы, то есть провести по крайней мере пять синхронных измерений. Ясно, что такая синхронизация возможна лишь при помощи магнитоэлектрического осциллографа. Отметки времени даются самим осциллографом. Два шлейфа проводят нулевые линии; один-два шлейфа нужно оставлять в качестве запасных. Поэтому осциллограф должен иметь не менее 7—8 шлейфов. Для измерения пройденного пути берем измерительное колесо, прижимаемое пружиной к поверхности следа одного из колес жатки; датчики контактные или индукционные; конструкцию измерительного колеса принимаем такой, при которой практически диаметр колеса не изменен, а буксование или проскальзывание отсутствуют.

Реохордный датчик на поводке рычажного устройства, в котором зазоры выбираются пружинами, позволит пяткой, скользящей сразу за колесом жатки, записывать расстояние между осью колеса и основанием, то есть прогибы шины. Такова же в принципе схема конструкции датчика, показывающего расстояние от середины ножевого бруса до почвы. Прогиб ножевого бруса измеряем при помощи мощных фольговых датчиков, наклеенных на середину бруса.

Отметим, что три последних датчика позволяют при достаточно большом количестве измерений приближенно вычислить статистические характеристики микрорельефа, что может при

необходимости служить контролем непосредственного и более точного измерения микрорельефа.

Из этого краткого описания подбора аппаратуры следует, что все записи осциллографом возможны без предварительного усиления.

Во втором примере имеется датчик, определяющий расстояние середины ножевого бруса от почвы. Заметим, что при достаточно большом количестве измерений показания остальных датчиков (за исключением датчика замера пройденного пути) в алгебраической сумме дают тот же результат. Поэтому определим как ошибку только что упомянутого первого датчика, так и суммарные показания остальных. Пусть первый датчик при калибровке показал, что предельная угловая ошибка качающегося рычага достигает $\pm 2\%$ при 90° отклонения рычага, что соответствует касанию бруса с почвой, или предельная относительная ошибка составляет менее $\pm 5\%$. Датчик прогиба бруса (табл. 1.12) дает предельную ошибку около $\pm 0,5\%$. Датчик прогиба шины при калибровке дал предельную ошибку около $\pm 5\%$.

Таким образом, на основании правила первого установления предельных относительных ошибок суммы в качестве общей предельной ошибки этой группы датчиков можно принять

$$\pm \frac{0,5 + 5}{2} \pm 2,7\%,$$

а общую ошибку записи осциллографом — около $\pm 5\%$.

Ошибки опытов. В соответствии с таблицей 1.10, класс второй, способы измерения можно разбить на два вида: измерения прямые, когда данную величину измеряют непосредственно, и измерения косвенные или посредственные, когда искомая величина является функцией измеряемых величин. Можно по-разному ставить опыт и выбирать способы измерений. Так как измерения косвенные зависят от ряда прямых, то при прочих равных условиях всегда выгоднее тот способ, при котором будет меньше прямых измерений, а значит, и меньше сумма ошибок.

Давая понятие «опыт» чисто количественное определение, можно сказать, что под опытом следует подразумевать совокупность разовых измерений различных величин и отметок в одних и тех же условиях.

Во многих исследованиях каждой точке кривой развития явления соответствует лишь один опыт; в этом случае каждое из качественно различных измерений проводят один раз и в данном опыте не повторяют. Если такой опыт состоит из одного прямого измерения, то

его ошибка равна ошибке одного измерения. Но как быть, если опыт состоит из нескольких качественно различных измерений — аргументов, каждое из которых дает какую-то ошибку? В этом случае точность опыта оценивают на основании вычисления предельной (наибольшей) ошибки опыта.

Из соображений, изложенных при рассмотрении формул (1.29) и (1.30), обычно вычисляют предельную относительную ошибку опыта на основании следующих правил.

1. Ошибка суммы заключена между наибольшей и наименьшей из относительных ошибок слагаемых; практически берут или наибольшую относительную ошибку, или среднюю арифметическую.

2. Ошибка произведения или частного от деления равна сумме относительных ошибок сомножителей или соответственно делимого и частного.

3. Ошибка n -й степени какого-либо основания (значения величины) в n раз больше относительной ошибки основания.

4. Ошибка синуса и косинуса равна произведению абсолютного значения котангенса измеренного угла или соответственно тангенса на предельную абсолютную ошибку угла, выраженную в радианах.

5. Ошибка тангенса и котангенса равна частному от деления двойной абсолютной ошибки угла (в радианах) на синус двойного угла.

Во всех случаях установления точности опыта точность вычислений результата должна определяться точностью измерений.

Если рассматривать ошибку измерения как частное значение переменной величины, предельную относительную ошибку опыта можно также вычислить по формуле:

$$\frac{\Delta_n(a)}{a} = \pm d |\ln f(u, v, w...)|, \quad (1.34)$$

где определяемая величина a является функцией переменных u, v, w , то есть предельная относительная ошибка величины a равна дифференциалу ее натурального логарифма, причем следует брать сумму абсолютных значений всех членов такого выражения.

Уравнение (1.34) дает возможность решить обратную задачу: определить необходимую точность измерений различными способами, если задана общая точность опыта. Но проще поступить иначе: установить заранее требования к метрологическим показателям приборов (цена или интервал деления, пределы измерения, порог чувствительности, измерительное усилие, погрешность и вариация показаний и т. д.), вычислить предельную ошибку и по ней подобрать недостающую аппаратуру и заменить приборы, дающие слишком

большую предельную погрешность, или части шкал приборов, их рабочие звенья.

Используя формулу (1.34), можно найти условия измерения, при которых предельная относительная ошибка функции будет наименьшей. Условием определенного решения здесь является наличие минимума функции. Найдя частные производные первого порядка по всем независимым переменным и приравняв полученные выражения нулю, вычисляют значения всех переменных. Определив знаки частных производных второго порядка при найденных значениях независимых переменных, получим дополнительные условия минимума.

Нижеследующий пример по определению лучших условий измерений заимствован у К. П. Яковлева.

При фотометрических работах поляризационный спектрофотометр на равную освещенность обеих половин поля прибора устанавливают поворотом окулярного николя. При этом связь между силой света I_0 некоторой длины волны в спектре эталонного источника света и силой света I_x той же длины волны в спектре исследуемого источника выражается формулой:

$$I_x = I_0 \operatorname{tg}^2 \alpha,$$

где α — угол, на который необходимо повернуть окулярный николь для установки равного освещения.

Требуется определить, при каких значениях угла относительная ошибка результата будет наименьшей.

Находим по формуле (1.34), третьему и пятому правилам величину относительной ошибки I_x :

$$\frac{\Delta_n(I_x)}{I_x} = \pm dI \ln I_0 \operatorname{tg}^2 \alpha = \pm 2d \ln \operatorname{tg} \alpha = \pm 2 \frac{2d\alpha}{\sin 2\alpha}.$$

Очевидно, что минимум относительной ошибки здесь соответствует максимуму $\sin 2\alpha$, отсюда $\alpha=45^\circ$.

Вывод. Ошибка будет наименьшей в тех случаях, когда окулярный николь необходимо поворачивать для измерений на углы, близкие к 45° . Ошибка возрастает, если по условиям опыта эти углы значительно отличаются от 45° .

Рассмотрим порядок вычисления предельной ошибки опыта в наших двух основных примерах. В первом примере установим предельную относительную ошибку вычислений производительности агрегата. Производительность агрегата (га в смену) может быть выражена так:

$$W_{\text{см}} = 0,1 B v T_p = 0,1 B \frac{S}{T_S} T_p,$$

где B — захват агрегата, м;

v — скорость, км/ч (прохождение пути S , км за время T_S , ч);

T_p — чистое рабочее время за смену, ч.

В этой формуле есть умножение и деление измеренных величин, поэтому в соответствии со вторым правилом

$$\frac{\Delta_n(W_{\text{см}})}{W_{\text{см}}} = \pm \left(\frac{\Delta_n(B)}{B} + \frac{\Delta_n(S)}{S} + \frac{\Delta_n(T_S)}{T_S} + \frac{\Delta_n(T_p)}{T_p} \right).$$

Учитывая подобранную в этом примере аппаратуру и средние данные таблицы 1.12, получим:

$$\frac{\Delta_n(W_{\text{см}})}{W_{\text{см}}} = \pm (0,25 + 0,25 + 0,55 + 0,55) \pm 1,6\%.$$

Из изложенного можно сделать вывод: для того чтобы правильно подобрать аппаратуру, надо провести сравнительную оценку точности различных способов измерения. В этом случае полезно также заранее задаться необходимой точностью опытов.

Точность измерений должна быть *целесообразной*. Можно указать на три основных обоснования целесообразности.

Первое заключается в практическом использовании результатов. Если, например, при технологическом процессе заточки лезвия нельзя добиться его затупленности меньше 10 мк, не следует измерить износ с точностью до десятых долей микрона.

Второе обоснование: в некоторых случаях нецелесообразно измерять с ошибкой, меньшей некоторых колебаний значений измеряемой величины. Например, может изменяться база отсчета, а по условию методики эти изменения не должны в известной мере приниматься во внимание. Таковы измерения глубины вспашки линейкой, точность которых может не превышать вспучивания почвы на краю борозды, обычно колеблющемся в пределах 5—10 мм, или диаметра отдельных комков на пути качения колеса измерительного прибора.

Третье обоснование касается экономической стороны. Чем точнее измерительная аппаратура, тем она сложнее и дороже, тем больше затраты на ее ремонт и калибровку. Может оказаться, что при малом количестве измерений и той же надежности, что и при большем количестве измерений прибором, дающим несколько большую ошибку, стоимость измерений дорогим прибором будет выше.

При определении величины случайных ошибок, кроме предельной ошибки, вычисляют *статистическую* ошибку неоднократных измерений. Эту ошибку устанавливают после измерений при помощи методов математической статистики и теории ошибок.

Если, например, диаметр вала, вязкость масла и т. д. измерять по одному разу, случайные ошибки могут сильно исказить результат. Поэтому лучше измерять какую-либо практически постоянную величину в тождественных условиях несколько раз и брать среднюю арифметическую этих измерений. Средняя арифметическая измерений является наиболее вероятным значением измеряемой величины при данном количестве намерений. В теории ошибок доказывается, что чем больше проведено измерений какой-либо величины, тем меньше становится суммарная случайная ошибка средней, и при бесконечном числе измерений случайная ошибка средней бесконечно мала.

Чем больше значение случайных ошибок и разброс (рассеяние) отсчетов, тем большее число раз надо измерить одну и ту же величину, чтобы достигнуть заданной точности или надежности измерений.

Рассеяние результатов измерений указывает на большую или меньшую их изменчивость и обычно оценивается средним квадратом отклонений наблюдаемых значений величины a_i , от их средних \bar{a} или квадратным корнем из среднего квадрата.

Средняя арифметическая \bar{a} равна сумме всех n отдельных результатов измерений a_1, a_2, \dots, a_n , деленной на число измерений:

$$\bar{a} = \frac{a_1 + a_2 + \dots + a_n}{n} = \frac{\sum_{i=1}^n a_i}{n}.$$

Если же все измерения сгруппированы в m классов с разными количествами измерений n_1, n_2, \dots, n_m в каждом классе, то следует вычислять взвешенную среднюю арифметическую:

$$\bar{a} = \frac{a_1 n_1 + a_2 n_2 + \dots + a_m n_m}{n_1 + n_2 + \dots + n_m} = \frac{\sum_{i=1}^m a_i n_i}{n}. \quad (1.35)$$

Здесь $\dot{a}_1, \dot{a}_2, \dots, \dot{a}_m$ — средние по классу.

Две последние формулы тождественны при $n_i=1$, так как тогда $m=n$.

Отклонение любого отдельного результата измерений от средней арифметической можно представить как разность

$$a_i - \bar{a},$$

где a_i — результат любого измерения.

Дисперсией случайной величины называется среднее значение квадрата отклонений случайной величины от ее среднего значения. Дисперсия обозначается через σ^2 и выражается формулой:

$$\sigma^2 = \frac{\sum_{i=1}^n (a_i - \bar{a})^2}{n - 1}.$$

Корень квадратный из дисперсии, взятый с положительным знаком, называется средним квадратическим отклонением, или стандартом:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (a_i - \bar{a})^2}{n - 1}}.$$

При разделении всех измерений на m классов с классовыми средними дисперсия будет равна:

$$\sigma^2 = \frac{\sum_{i=1}^m n_i (a_i - \bar{a})^2}{n - 1},$$

а стандарт

$$\sigma = \sqrt{\frac{\sum_{i=1}^m n_i (a_i - \bar{a})^2}{n - 1}}. \tag{1.36}$$

Стандарт имеет размерность величины, для которой он вычислен.

Дисперсия и стандарт — это меры рассеяния, или изменчивости; чем больше дисперсия или стандарт, тем больше рассеяны значения измерений.

Таким образом, при измерении неизменной величины среднее квадратическое отклонение (стандарт) является мерой точности среднего арифметического значения неоднократно измеренной неизменной величины.

Если же неоднократно измеряемая величина переменна, то вычисленное по ее измерениям значение стандарта показывает не только меру точности как случайную ошибку измерений, но и меру изменчивости переменной.

Остальные встречающиеся в литературе понятия точности (средняя ошибка, вероятная ошибка) не имеют преимуществ перед стандартом, являются его производными и поэтому нами рассматриваться в этом разделе не будут.

Абсолютное значение стандарта зависит и от совершенства измерительных приборов. Если одну и ту же величину измерять при помощи приборов различной точности, абсолютное значение стандарта будет меньше при измерении более точным прибором. Например, если диаметр отверстия измерять сначала нутромером со шкалой в миллиметрах, а затем индикатором со шкалой в микронах, значение стандарта при последних измерениях будем меньшим. Это обстоятельство имеет важное значение при установлении числа измерений или опытов.

Для подавляющего большинства технических измерений можно считать, что наибольшей ошибкой средней арифметической многократных измерений является абсолютная величина, равная плюс — минус трем стандартам,

$$\Delta_n(\bar{a}) = \pm 3\sigma \quad (1.37)$$

или относительная величина

$$\frac{\Delta_n(\bar{a})}{\bar{a}} = \pm \frac{3\sigma}{\bar{a}}; \quad \frac{\Delta_n(\bar{a})}{\bar{a}} 100 = \pm \frac{3\sigma}{\bar{a}} 100\%. \quad (1.38)$$

Эту ошибку назовем наибольшей возможной статистической в отличие от предельной ошибки, характеризуемой формулой (1.33).

Здесь в качестве примера приведем результаты проверки на точность показаний пневматического калибратора, которым измеряли относительную неплотность в одном из цилиндров тракторного дизеля (данные Л. Х. Морозова, Б. И. Иванова, В. М. Славущкого, Н. М. Рогожкина и Л. А. Зайцева). Было проведено 20 измерений, между которыми двигатель работал по 2—3 мни на оборотах, близких к номинальным, и получен ряд значений неплотностей (мм²): 0,465; 0,450; 0,450; (1,150; 0,465; 0,465; 0,465; 0,450; 0,435; 0,465; 0,450; 0,465; (1,165; 0,435; 0,435; 0,450; 0,465; 0,465; 0,450.

Средняя

$$\bar{a} = \frac{9,185}{20} \approx 0,459 \text{ мм}^2.$$

Стандарт

$$\sigma = \sqrt{\frac{0,00265}{20-1}} \approx 0,012 \text{ мм}^2.$$

Отсюда по формуле (1.37) наибольшая возможная статистическая ошибка

$$\Delta_n(\bar{a}) = \pm 0,012 \cdot 3 = \pm 0,036 \text{ мм}^3$$

или (по формуле 1.38)

$$\frac{\Delta_n(\bar{a})}{\bar{a}} 100 = \pm \frac{0,036}{0,459} 100 \approx \pm 7,8 \%$$

Если предельную ошибку устанавливают до измерений, то наибольшую статистическую ошибку вычисляют по результатам неоднократных измерений. Обычно наибольшая статистическая ошибка при измерении неизменной величины меньше предельной, так как отклонения отдельных измерений от средней не однозначны, как это принято для предельной ошибки. Иногда и значения измеряемой величины отсчитывают с большей точностью, чем это предположено для предельной ошибки. Если же измерения проводить небрежно, статистическая ошибка может быть больше предельной.

Точность разовых измерений оценивают только по предельной ошибке. При неоднократных измерениях до их начала следует пользоваться предельной ошибкой (для подбора аппаратуры и представления о возможностях измерений), а после измерений оценивать их точность по наибольшей возможной статистической ошибке.

Предельную ошибку косвенных измерений при пользовании наибольшими статистическими ошибками прямых измерений находят по формуле (1.34) и правилам изложенным выше.

Измерения, дающие дисперсии или стандарты одинаковой величины, называются равноточными. Равноточность измерений серии опытов облегчает обработку результатов наблюдений и уменьшает суммарные ошибки исследования. Практически добиться равноточности измерений можно только в тех случаях, когда измерения будут проводить достаточно опытный человек одним и тем же прибором в одинаковых условиях.

В некоторых случаях неравноточность приводит к несравнимым результатам, к ошибочным выводам и заключениям, поэтому надо добиваться равноточности измерений данной величины. В дальнейшем мы будем исходить преимущественно из предположения о равноточности измерений.

1.7.3. Наблюдение

Пассивное наблюдение (без вмешательства в условия развития явления) непосредственно целям экспериментального исследования служить не может. Связывая пассивное наблюдение с индуктивным

методом, В. П. Горячкин писал: «Этот способ неизбежен, когда приходится иметь дело с процессами, не поддающимися какому-либо анализу, и является наиболее распространенным, но наименее плодотворным».

Пассивное наблюдение в экспериментальном исследовании применяют преимущественно для предварительной проверки общей правильности рабочей гипотезы, то есть для установления направленности развития явления (например, изучение изменения формы деталей в процессе естественного изнашивания) и факторов, обуславливающих развитие явления. Такое наблюдение используют также при изучении и численном определении некоторых свойств исследуемого объекта (например, размеров зерна, его парусности и т. д. при исследовании сортирования зерна) и при изучении условий развития явления (например, изучение почвы, рельефа и формы участков, отведенных под эксперимент).

При пассивном наблюдении исследователь регистрирует различные интересующие его стороны развития явления. Для регистрации употребляют разнообразные средства измерений и отметок, вплоть до самых совершенных. Пассивное наблюдение можно чередовать (во времени) с активным (экспериментом, опытом).

Наблюдение становится активным, когда исследователь сам определяет условия развития явления, когда он видоизменяет эти условия в желаемом направлении, чтобы получить ясные закономерности. Такое наблюдение наиболее плодотворно.

1.7.4. Поисковые опыты

Первой ступенью активного наблюдения являются поисковые опыты. Если наблюдение пассивное играет в исследовании роль «разведки наблюдением», то поисковые опыты представляют собой «разведку боем».

Поисковые опыты проводят не столько для раскрытия закономерностей развития явления, сколько в тех же целях, что и пассивное наблюдение, если последнее вместе с изучением предшествовавших исследований не раскрывает условий развития и направленности явления, а также обуславливающих его факторов. В отличие от пассивного наблюдения поисковые опыты ставят в заранее назначенных условиях. Возьмем первый пример. Анализ литературы и пассивное наблюдение дали общую схему развития явления, но не смогли ясно ответить на вопрос: будут ли по мере затупления лемехов изменяться отклонения от средней глубины пахоты, и если да, то в какую сторону? Это обстоятельство может быть выяснено поисковыми опытами, если оно играет достаточно важную роль в исследовании.

Другая задача поисковых опытов — проверить варианты рабочей гипотезы, если на одном трудно остановиться, а провести исследование по двум вариантам невозможно.

Далее, задачей поисковых опытов может быть установление факторов, определяющих развитие явления или отбор основных факторов (см. пункт «Основные опыты»).

Целью поисковых опытов может быть и проверка отдельных частей разработанной методики и приспособленности приборов к тем измерениям, которые определила методика.

Поисковые опыты можно ставить и для того, чтобы найти основания для расчета числа опытов.

Таким образом, поисковые опыты могут предшествовать разработке общей методики, но могут быть и включены в нее (желательно особым разделом). В некоторых случаях целесообразно не устанавливать рабочей гипотезы и не составлять общей методики до получения результатов по поисковым опытам.

Чтобы провести поисковые опыты, необходимо составить частную методику. Число поисковых опытов можно назначить на основании следующих предположений.

1. Если проверяют направление процесса, обычно достаточно двух опытов — в начале и в конце процесса.

2. Если устанавливают, какие факторы обуславливают развитие явления, то наименьшее число поисковых опытов примерно равно удвоенному предполагаемому числу факторов, так как для того, чтобы выяснить влияет ли данный фактор на развитие явления или нет, необходимо поставить как минимум два опыта — в начале (когда данный фактор явно отсутствует) и в конце (когда данный фактор явно присутствует) развития явления или в какой-то точке области оптимума.

3. При проверке вариантов рабочей гипотезы выбирают главный, стержневой фактор гипотезы и по небольшой серии, 3—5 опытов, определяют ход изменения основных кривых рабочей гипотезы. Например, в первом примере осталось неясным, будут ли по мере затупления лемеха изменяться отклонения от средней глубины вспашки, и если да, то в какую сторону? Для уточнения этого после установки всех приборов на агрегат и их опробования проводим два поисковых опыта вспашки какой-либо почвы: один с новыми лемехами, другой — с лемехами, весьма затупленными. Результаты показывают, что затупленные лемеха дают большие отклонения от средней глубины вспашки, чем новые.

Кроме того, полученные в этих поисковых опытах динамограммы дают возможность обосновать для данных условий эксперимента количество измерений.

По второму примеру неясна практическая зависимость колебаний бруса от скорости. Будет ли размах колебаний находиться в степенной зависимости от скорости? Очевидно, что для этого надо после опробования установленных на машины приборов провести три поисковых опыта: при наименьшей скорости движения, при наибольшей и, наконец, при средней скорости движения.

Не указывая на детали поисковых опытов (они станут более ясными при изучении требований к проведению основных опытов), будем считать, что поисковые опыты дали именно степенную зависимость потерь в функции скорости. Но бывают случаи, когда поиск превращается в самостоятельное исследование.

Практически исследователь должен исходить из следующих условий: ***факторы, которые могут определять развитие явления, известны, но их действие как по значимости, так и по направлению неизвестно; конкретная числовая цель исследования или результат, которого следует добиться при помощи той или иной комбинации известных факторов, известны; обработка опытов нетрудоемка, не отнимает много времени.***

Главное здесь — найти методику, осуществлением которой цель была бы достигнута (конечно, с условием, что она достижима при том или ином сочетании уже известных факторов) при наименьшем числе опытов.

Идея такой методики заключается в том, что строится векторная диаграмма опытов, для построения которой результаты каждого опыта (или каждой серии опытов) по сочетанию факторов обсуждаются, и для следующего опыта (серии опытов) выбирается предположительно такое сочетание факторов, результаты которого давали бы вектор, лучшим образом ведущий к цели. Диаграмма в зависимости от условий может быть плоскостной или пространственной или представлять собой комбинации плоских или пространственных диаграмм. Первый опыт (первая серия опытов) начинается при любом таком сочетании факторов, которое исследователь считает наиболее логичным.

В качестве простого примера возьмем следующий. Пусть имеем присадку к приработочному (обкаточному) маслу. Требуется определить оптимальное содержание присадки, дающее заданную шероховатость поверхности прирабатываемых деталей в наименьшее время приработки, причем известно, что содержание

присадки в масле не должно превышать одного процента по весу и что на каждую единицу времени опытов с данным содержанием присадки изменение шероховатости равномерно.

Если задачу и ее решение изобразить в системе прямоугольных координат времени (t) и шероховатости (y), то заданная шероховатость (y_3) может быть представлена прерывистой линией, параллельной оси t (рис. 1.16).

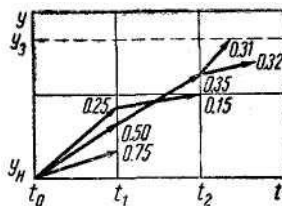


Рис. 1.16. Схема простого опыта движения по градиенту.

Через y_H на рисунке обозначена начальная шероховатость.

Ось времени приработки разбита на равномерные, но достаточно малые (по опыту предыдущих работ), участки t_1, t_2, t_3, \dots , по происшествии которых измеряются шероховатости. Естественно предположить, что в самом начале за период $t_0—t_1$ лучше поставить три опыта с 0,5; 0,75; 0,25% присадки (конечно, в первой серии может быть и больше и меньше опытов и различие в присадке может быть иным). Полученное обозначаем векторами, на конце которых находятся точки результатов опытов. Как видно, и этом случае модули векторов здесь прямо связаны со скоростью приближения к заданной неровности. Обсуждая результат, заключаем, что дальнейшие опыты с содержанием присадки $>0,5\%$ можно опустить. Таким образом, дальнейший опыт следует провести с присадкой $<0,5\%$ и $<0,25\%$ (чтобы сравнить их модули за отрезок $t_1—t_2$). Предположим, что для исследования влияния содержания присадки $<0,25\%$ назначим присадку в 0,15%, а для исследования влияния разницы между 0,25 и 0,50% назначим содержание присадки в 0,35%. За время t_2 производим измерения шероховатостей и вычерчиваем соответствующие векторы. Сопоставляя модули всех векторов, приходим к выводу, что оптимум содержания присадки находится между 0,25 и 0,35% прибавки, причем полагаем, что этот оптимум должен быть несколько выше

$$\frac{0,25 + 0,35}{2} = 0,30\%.$$

Назначаем 0,31 и 0,32%, а так как векторы, очевидно, покажут большую скорость приближения к заданному пределу, время опытов

сокращаем. Проведя замеры, получаем два новых вектора, из модулей которых следует, что оптимальным количеством присадки можно считать 0,31%.

Как видно, чем круче положение вектора, тем меньше может быть интервал или шаг опытов. Таким образом, за семь шагов опытов типа поисковых найдено решение задачи. Внимательный читатель уже заметил, что весь поиск свелся к простейшему нахождению направления наибыстрейшего возрастания функции (направления градиента).

Методы поиска, основанные на анализе производных градиентов и применении некоторых приемов теории вероятностей, при помощи которых можно получить заданный результат с минимальным числом опытов, будут изложены в нашей работе дальше. Этот метод называется движением по градиенту или крутым восхождением.

1.7.5. Основные опыты

При разработке рабочей гипотезы устанавливают факторы, определяющие развитие явления. Иначе говоря, устанавливают независимые переменные (аргументы), определяющие изменения зависимых переменных (функций).

Исследователь, наблюдая в опыте за развитием явления при воздействии факторов, получает материал или данные исследования. В опытах измеряют или отмечают величины и качественные показатели, характеризующие как факторы (аргументы), так и показатели развития явления (отклики, функции).

Обычно в первую очередь определяют величины и свойства, характеризующие факторы. Примеры: при исследовании способов выделения из общей массы семян, пораженных суховеем, предварительно изучают биологические и физико-механические свойства семян, подлежащих разделению; когда исследуют пахоту, изучают и отбирают типы почв и их свойства; если изучают энергетические показатели и качество работы силосоуборочных комбайнов, определяют густоту стояния и высоту растений, а также их физико-механические свойства.

Множественность величин, подлежащих измерению, может по тем или иным соображениям затруднять исследование.

Изучение суммарного влияния многих количественно не определенных факторов, сочетающихся в произвольных и непонятных для исследователя соотношениях, часто приводит к неясным закономерностям и ошибочным выводам. *В таких случаях следует все факторы, обуславливающие явление, разделить на **основные**, оказывающие наибольшее влияние на развитие явления, несущие*

наибольшую информацию, и дополнительные, влияющие на развитие явления второстепенно. Тогда в опыте измеряют или отмечают лишь величины и свойства, характеризующие основные факторы.

Выделение основных факторов, как главных звеньев цепи исследования, весьма важно, но следует также учитывать условность деления факторов на основные и дополнительные, так как даже при небольшом изменении условий опыта дополнительные факторы могут стать основными и наоборот. Кроме того, развитие науки может показать, что факторы, ранее считавшиеся основными, составлены из ряда дополнительных, и поэтому те из них, которые ранее были дополнительными, развитием познания могут стать основными.

Далее следует всегда помнить, что, если не учитывать дополнительные факторы, можно допустить серьезную ошибку, так как количественные изменения, вносимые дополнительными факторами, даже если последние правильно определены, могут привести к качественным изменениям процесса. Так, например, при исследовании процесса сгорания в цилиндрах двигателя одним из главных факторов считалась степень распыления топлива, а состояние пленки последнего на поверхности камеры сгорания — второстепенным. Однако точное исследование показало, что именно учет пленочного состояния и паровой фазы топлива, дает возможность использовать его лучшим образом.

В. П. Горячкин пришел в свое время к выводу: «... Выбор основных главных элементов исследования. Эта задача является наиболее трудной, и от удачи выбора элементов зависит успех исследования. Надо помнить, что всякое явление чрезвычайно сложно и определяется множеством элементов, из которых надо выбрать только необходимые и достаточные».

Чтобы устранить или по крайней мере уменьшить ошибку, появляющуюся вследствие деления факторов на основные и дополнительные, *надо при постановке опытов стремиться нейтрализовать дополнительные факторы, то есть создать такие условия, при которых действие дополнительных факторов было бы возможно более неизменно и незначительно, а величины и свойства, характеризующие эти факторы, приближались бы к постоянным.* При этом исследователь должен стремиться сделать переменными величинами лишь **основные факторы**. Таким образом, **общим принципом исследования является постоянство всех остальных факторов при изменении избранных.**

Есть четыре основных приема нейтрализации дополнительных факторов.

Метод резкого изменения переменных факторов при относительно малом изменении остальных. Сопротивление плуга (первый пример) измеряют для сравнительно резко изменяющегося затупления лемеха, при этом в опытах используют один участок почвы с постоянной влажностью, а плуг пашет на одну и ту же среднюю глубину с одинаковой скоростью движения и т. д. Таким образом находится закономерное изменение сопротивления плуга в функции затупления лемеха.

Во втором примере по определенному прямому пути движется один и тот же агрегат с одинаковым давлением в новых или равно изношенных шинах и преодолевает одинаковые по форме, но разные по высоте препятствия. Здесь регистрируют затраты мощности или крутящего момента в функции высоты препятствий.

Если исследуют изменение физико-механических свойств растения в функции времени (степень зрелости), растения необходимо брать с одних и тех же участков с однородными климатическими условиями, отводя для отбора растений площадки с возможно более выравненными растениями, так как и противном случае могут получиться несовместимые результаты.

Метод контрольных опытов, когда быстро и в любой мере меняющиеся дополнительные факторы одновременно воздействуют на несколько градаций основных факторов, одну из которых считают контрольным эталоном и с ней сравнивают все остальные. При исследовании изнашиваемости лемехов плуга, новых или отремонтированных различными способами, на один и тот же плуг ставят два вида лемехов. Для наибольшей нейтрализации остальных факторов каждый вид лемехов монтируют через один корпус. Тогда быстро меняющиеся факторы (скорость, состояние и состав почвы и др.) действуют на лемеха различного качества практически одновременно, но не в одинаковой степени, так что в сравнимых условиях выясняется износостойкость различных лемехов.

Контрольные опыты применяют также при сравнительной проверке влияния различных технологических процессов, чтобы выявить лучший из них. Для этого выбирают поле с возможно более однородным почвенным покровом и одинаковыми условиями увлажнения и разделяют его на количество участков, кратное количеству сравниваемых технологических процессов. Эти участки выравнивают биологически и технически, то есть минимум в течение одного года одинаково обрабатывают и удобряют, высевают одинаковыми приемами одну культуру и убирают ее одним и тем же способом. Так как почва участков, несмотря на эти мероприятия, все

же неоднородна и является в той или иной мере комплексной, назначают число участков контрольных и для проверки новых технологических процессов и располагают их так, чтобы добиться сравнимости результатов и достаточной надежности.

Этих комбинаций может быть много, и число их связано прежде всего с необходимостью нейтрализовать случайные изменения почвы. Поэтому большие участки разбивают на опытные делянки, причем контрольные делянки перемешиваются с делянками для проверки новых технологических процессов. Принципиального различия в разбивке участков и делянок для этой цели по сравнению с методами, применяемыми в земледелии, агрохимии и растениеводстве, здесь нет.

Метод «чистых» опытов. При этом методе стремятся искусственно создать условия, в которых дополнительные факторы не проявлялись или не влияли бы при проведении опытов на изменяющиеся основные факторы. Этот метод используют лишь в лабораторных опытах. Например, чтобы исследовать действие различных деформаторов на почву различного состава, заменяют все многообразие почв искусственной однородной почвой вполне определенных свойств. «Чистые» опыты хорошо выявляют закономерности, но обычно требуется синтез ряда таких «чистых» закономерностей, чтобы выводы из опытов имели производственное значение.

Метод разных знаков состоит в том, что одному и тому же фактору в опытах придают сначала положительное значение, а затем отрицательное (или наоборот), чтобы при вычислении средней взаимно погасились ошибки от неучета влияния этого фактора. Этот метод применяют прежде всего при исследовании механики явлений и гораздо реже — в исследованиях технологического процесса. Например, при исследовании сопротивления плуга (первый пример) или дополнительной силы сопротивления (второй пример) некоторый, хотя бы и малый, наклон участка, на котором проводят испытания, может привести к ощутимой ошибке. Чтобы погасить эту ошибку, обычно измеряют величины при ходе в одну (туда), затем при следующем, рядом расположенном, ходе в другую сторону (обратно).

Возникает вопрос: сколько же основных факторов принять для измерений и отметок в опыте? На этот вопрос нельзя ответить однозначно. Здесь прежде всего следует на основании практики предыдущих исследований или поисковых опытов установить, можно ли разделить прямыми измерениями влияние переменных факторов на развитие явления, то есть имеются ли в распоряжении исследователя способы непосредственного, в опытах, установления влияния того или иного переменного фактора.

Если эти условия есть, при таких измерениях необходимо проследить влияние всех возможных факторов на развитие явления.

Многофакторные исследования всегда дают более полную и более жизненную картину синтеза различных закономерностей.

Например, исследуют влияние изнашивания отдельных деталей на процесс изнашивания двигателя внутреннего сгорания в целом. Метод радиоактивных изотопов позволяет (при использовании изотопов с различной проникающей способностью или различной энергией излучения) выделить 2—3 группы деталей, изнашивание которых регистрируется одновременно.

Но возможность одновременно селективно изучать влияние ряда факторов имеется не всегда, и поэтому, чтобы получить ясные закономерности с достаточной степенью надежности, во многих случаях следует выделить в качестве основного фактора лишь один (однофакторный опыт) и нейтрализовать все остальные факторы. *Изменение одного фактора при постоянстве остальных—характерная черта многих надежных исследований.*

Есть математические методы, при помощи которых можно численно оценить влияние каждого из факторов при многофакторном исследовании без селективного одновременного измерения величин по отдельным факторам. Но трудоемкость реализации этих методов велика. Она увеличивается с увеличением количества одновременно влияющих основных факторов, и уже при четырех факторах выделить влияние каждого из них математическими методами трудно. В этих случаях, если необходимо установить закономерности влияния нескольких факторов, приходится расширять исследование, увеличивая количество измерений и разделяя опыты на серии, в каждой из которых и ряде однотипных опытов проводят однофакторное (при возможности двухфакторное) исследование. Все это увеличивает трудоемкость исследования и затраты времени, что следует учитывать.

После того как установят основные факторы и способы нейтрализации величин, влияние которых желательно оставить практически неизменным, выбирают способы, направления и пределы изменения основных факторов.

Во многих случаях все это входит в цели исследования. Если, например, надо изучить влияние скорости передвижения агрегата, составленного из определенных машин, на технологический процесс уборки, то способ, направление и предел изменения скорости, как основного фактора, задан конструкцией трактора или уборочной машины.

Решение становится более сложным, если способ, направление или пределы изменения не заданы условиями и их необходимо назначить. В этих случаях решение целесообразно принять исходя из следующих соображений. Если определяют способ изменения основного фактора, руководствуются принципом подобия естественным условиям. Например, при изучении влияния износа детали на показатели работы машин ее форма может быть изменена искусственно, но с учетом изменения формы при естественном износе. Когда устанавливают пределы изменения факторов, учитывая практику, несколько расширяют практически принимаемые максимальные и минимальные (а не средние) пределы.

Направление изменения факторов — это, скорее, организационная или экономическая сторона исследования. Здесь надо выбрать: проводить ли поисковые или поставить основные опыты во всех возможных направлениях.

При разработке методики следует стремиться получить в опытах данные, которые не только устанавливают закономерности развития явлений, но и помогают найти объяснение этих закономерностей, их причины, физическую основу явления. Чем полнее выяснена физическая основа явления, тем значительнее исследование, тем практичнее и жизненнее его результаты.

Чем полнее, тщательнее и продуманнее составлена методика исследования, тем легче и скорее выполнить исследование; ясность, точность и достаточно подробное (но краткое) изложение особенно важны в тех случаях, когда в исследовании принимает участие группа работников.

Продолжая первый пример, приходим к выводу, что, если нет синхронной записи, необходимо проводить однофакторное исследование. Из литературы и наблюдений знаем, что на сопротивление плуга, а следовательно, на скорость пахотного агрегата и расход топлива влияют следующие основные факторы: степень затупления лемеха, тип и влажность почвы, глубина вспашки и ширина захвата плуга. В соответствии с целями исследования за переменную величину принимаем затупление лемехов. Следовательно, надо провести серии опытов: на нескольких типах почв (столько серий, сколько типов почвы, разных по влажности), при различной глубине и ширине захвата (столько серий, сколько разностей по глубине и захвату).

Во втором примере основными факторами являются: высота, форма и свойства препятствия (упругость, сминаемость и т.д.), скорость данного агрегата и его упругие свойства. Из этих факторов синхронно

могут быть записаны последние два. Все факторы, значение которых не может быть записано синхронно, и модификации этих факторов определяют количество серий

1.7.6. Определение общего количества опытов

Экспериментальное исследование может состоять из одной или нескольких серий опытов. Сколько же опытов надо провести в исследовании?

Следует руководствоваться двоякого рода положениями, определяющими общее количество опытов (как сумм разовых измерений и отметок) в каждой серии.

Во-первых, необходимо такое количество опытов, которое достаточно выявило бы кривые функциональных зависимостей; иначе говоря, следует определить такое минимальное количество опытных точек, которое хорошо представляло бы кривую, поэтому предполагаемый характер кривых функциональных зависимостей фактически и определяет количество опытов.

Положение прямой вполне определено двумя точками. Если исследователь уверен в том, что данная зависимость прямая, он назначает два опыта с достаточным промежутком между ними (близко к пределам), чтобы увеличить точность положения прямой. Дуга, описанная одним радиусом, вполне определяется тремя точками, которые следует расположить так: две близко к пределам и одну в середине, чтобы точнее найти положение кривой.

Если кривые более сложные, такого количества опытов недостаточно. Для таких кривых применяют следующее правило: *рассматривая сложную кривую как комбинацию прямых и простых кривых, описать каждый перегиб кривой по меньшей мере тремя опытами; каждый участок, близкий к прямолинейному, двумя опытами и близко к назначенным пределам поставить два «концевых» опыта.*

Чтобы установить количество опытов, следует использовать графики закономерностей рабочей гипотезы. В каждом из графиков выбирают наиболее сложную кривую и по ней определяют количество опытов. В первом примере (рис. 1.3) есть две сложные кривые — средней глубины вспашки и сопротивления плуга. Обе кривые имеют по два перегиба и по одному участку, близкому к прямой. Таким образом, общее количество опытов будет равно: $1+3+2+3+1 = 10$. Если учесть, что концевые точки входят в количество описывающих перегибы кривых, то число опытов можно сократить до 8.

Если нужно заранее наметить расположение изменений величины основного фактора, то есть опытных точек, установленное количество опытных точек располагают по линии абсцисс с равномерными промежутками, так как действительные места изгибов кривых не известны. Такое распределение опытов уточняют во время исследования.

Если требуется установить не только общие закономерности, но и возможно более точно численные значения функций, каждый перегиб кривой должен быть обоснован минимум пятью опытами. К такого рода кривым относятся, например, кривые характеристик автотракторных двигателей, так как по их численным значениям оценивают конструкции.

Кроме того, особо тщательно должны изучаться резкие перегибы кривых или скачкообразные изменения развития явления, так как эти перегибы или скачки могут характеризовать переход количественных изменений в качественные и обычно бывает важно наиболее точно определить момент такого перехода.

Во-вторых, необходимо считаться со случайными ошибками опыта. Для уменьшения значения таких ошибок опыты иногда целесообразно повторять и брать среднюю арифметическую. Каково же должно быть количество повторностей? Из теории ошибок известно, что количество опытов (или измерений одной и той же величины) зависит от стандарта измерений и заданной надежности результата. Здесь под надежностью подразумевается вероятность получения тех же результатов при новых измерениях этой же величины или вероятность тех же результатов при повторении опыта в аналогичных условиях.

Чем больше относительные колебания результатов и чем большую надежность опыта желательно получить, тем больше должно быть повторностей.

Наиболее строго (для практического пользования) эта зависимость установлена В. И. Романовским. В таблице 1.13 показана зависимость между необходимым числом повторностей опыта (измерения) и заданными надежностью и относительной ошибкой.

Для того чтобы найти по этой таблице необходимое количество опытов (измерений одной и той же величины), необходимо задаться надежностью N и ошибкой Δ , взятой в долях стандарта σ .

Таблица 1.13.

Ошиб- ка Δ	Надежность опыта H							
	0,5	0,6	0,7	0,8	0,9	0,95	0,99	0,999
3,0	1	1	1	1	2	3	4	5
2,0	1	1	1	2	3	4	5	7
1,0	2	2	3	4	5	7	11	17
0,5	3	4	6	9	13	18	31	50
0,4	4	6	8	12	19	27	46	74
0,3	6	9	13	20	32	46	78	127
0,2	13	19	29	43	70	99	171	277
0,1	47	72	169	266	273	387	668	1 089
0,05	183	285	431	659	1 084	1 540	2 659	4 338
0,01	4543	7090	10 732	16 436	27 161	38 416	66 358	108 307

Как было выяснено, абсолютное значение стандарта зависит от точности измерительного прибора, поэтому надежность измерений зависит как от точности измерительного прибора, так и от количества измерений. *Чтобы добиться желаемой надежности, почти всегда выгоднее увеличить точность прибора, а не количество измерений.* Пусть например, при измерении менее точным инструментом какого-либо размера стандарт равен 0,90 мм, и более точным — 0,15 мм. Пусть допустимая ошибка измерений при надежности 0,95 должна быть не более 0,30 мм, то есть 1/3 стандарта при измерении менее точным и 2 стандарта при измерении более точным инструментом. По таблице 1.13 узнаем, что в этих условиях требуется более 27 измерений менее точным и только 4 измерения более точным инструментом.

Какой же надежностью следует задаваться в вычислениях? Строгих правил в выборе надежности нет. На практике при исследованиях закономерностей в самом общем виде (без деталей), например характера кривых развития явления, достаточна надежность $H=0,67$, или по таблице 1.13 $H=0,7$. Для измерений, связанных с конструкциями систем, вполне достаточна надежность $H=0,9$. При определении деталей закономерностей и значений величин, являющихся основой для дальнейшего расчета, необходима надежность $H = 0,99$.

Если поисковые опыты не проводились и нет данных, по которым можно было бы установить значение среднего квадратического отклонения (стандарта) σ , можно принять, что предельная ошибка приближенно равна наибольшей возможной статистической, то есть, принимая во внимание формулы (1.31—1.33) и (1.37—1.38),

$$\Delta_n \approx \pm 3\sigma.$$

Приведем примеры пользования таблицей 1.13. Пусть следует измерить микрометром (при отсутствии систематической ошибки или при внесении соответствующей поправки) какую-либо величину. Примем на основании формулы (1.37) предельную ошибку $\Delta_n = \pm 0,01$ мм. Тогда одно измерение микрометром по таблице 1.13 (пересечение данных строки $\Delta_n = 3\sigma$ соответствующими столбцами) дает надежность 0,8, два измерения – 0,9, три – 0,95. Последнее можно прочесть так: результаты трех измерений повторятся в 95 случаях из 100, если принять предельную ошибку микрометра $\pm 0,01$ мм. Можно также сказать, что, если требуется, чтобы в 95 случаях из 100 ошибка измерения микрометром не превышала $\pm 0,01$ мм надо сделать три измерения, взять среднюю измерений, округлив результат до сотых долей миллиметра.

При измерении переменных величин таблицу 1.13 также можно использовать для установления наиболее вероятной ошибки средней колеблющейся величины при заданной надежности, если известны средняя, количество измерений и стандарт. Пусть 250 раз измерено тяговое сопротивление машины и вычислено среднее сопротивление 502 кг и стандарт 63 кг. Зададимся надежностью 0,90. Берем ближайшее к 250 число измерений в этом столбце, то есть 273, и получаем ошибку средней 0,10, или $0,1 \cdot 63 = 6,3$ кг. Таким образом, действительное значение среднего тягового сопротивления машины находится в пределах $502 \pm 6,3$ кг. При необходимости более точные данные из таблицы 1.13 получают интерполированием.

Рассмотрим более сложный случай, связанный с заданной допустимой ошибкой вычисления среднего арифметического значения измеряемой величины. Иначе говоря, выясним, как при помощи таблицы 1.13 установиться сколько опытов или измерений надо выполнить при избранных нами точности и надежности вычисления средней. Пусть необходимо определить путь (расстояние), который должен пройти пахотный агрегат при измерениях в первом примере. В поисковых опытах выяснилось, что сопротивление плуга колеблется и запись сопротивления на ленте динамографа дает зубчатую кривую с колебаниями различной частоты и амплитуды. *Как правило, в подобных случаях для дальнейших вычислений следует брать наиболее типичные по амплитуде и периоду колебания.*

Тип колебаний выбирают в зависимости от цели исследования. При вспашке, например, может быть несколько вариантов выбора типичных колебаний. Во-первых, целью исследования могло быть определение влияния конструкции ходового устройства трактора на динамику плуга; в этом случае исследователь ищет характерные

колебания, связанные, например, с перематыванием гусениц. Во-вторых, если по цели исследования обработка комплексных почв должна совершаться без переключения передач, типичными будут колебания при определении пятен почвы с высоким сопротивлением.

Наконец, цели исследования могут быть связаны с отысканием максимальных амплитуд, соответствующих каким-то условиям (например, проверка на прочность). Допустим, исследовалась вспашка культурных почв. В этих условиях в поисковых опытах 33 типичных колебания сопротивления плуга были получены на отрезке ленты динамографа, соответствующем 5 м пройденного пути; стандарт оказался равным $\sigma_d=363$ кг, средняя сопротивлению $R = 1490$ кг. Какой должна быть минимальная длина пути плужного агрегата при работе в этих условиях, чтобы ошибка средней сопротивлению не превышала 2% при надежности 0,99?

Прежде всего отделим рассеяние точек, зависящее от взаимодействия «плуг — почва», от рассеяния, вызываемого ошибкой прибора. При калибровке динамометра получаем $\sigma_n = 18$ кг.

Тогда в соответствии с формулой

$$\sigma_0 = \sqrt{\sigma_d^2 - \sigma_n^2},$$

(где σ_d — стандарт двойного рассеяния) искомое рассеяние будет равно:

$$\sigma_0 = \sqrt{363^2 - 18^2} = 362 \text{ кг,}$$

а абсолютная ошибка средней: $1490 \cdot 0,02 = 30$ кг.

Вычислим эту ошибку в долях стандарта:

$$\Delta = \frac{30}{362} = 0,083.$$

По таблице 1.13 и ближайшей величине $\Delta=0,1$ находим, что необходимое количество колебаний сопротивления равно 668. Принимая, что на единицу длины ленты динамографа количество колебаний остается более или менее постоянным, получим минимальную протяженность пути, обеспечивающую результаты с надежностью 0,99 при ошибке средней около 2%:

$$\frac{668}{33} \cdot 5 = 101 \text{ м.}$$

Но в опытах этого примера нас интересует не только сопротивление, но и глубина вспашки. Условимся, что в поисковом опыте на протяжении 20 м пути измерения глубины вспашки показали 42 отклонения при средней $\bar{a} = 21$ см и при стандарте $\sigma = \pm 3$ см. Пусть ошибка средней не превышает 2% при надежности 0,99. Тогда абсолютная ошибка средней должна быть не более $21 \cdot 0,02 = 0,42$ см, а взятая

в долях стандарта $\Delta = 0,42: 3 = 0,14$. По таблице 1.135 находим, что при $\Delta' = 0,1$ $n' = 668$ и при $\Delta'' = 0,2$ $n'' = 171$. Упрощая расчеты и не интерполируя, находим среднее количество отклонений:

$$a = \frac{n' + n''}{2} \approx 420.$$

Тогда минимальная протяженность пути для опыта при условии, что частота отклонений на единицу пути изменялась мало, будет равна:

$$\frac{420 \cdot 20}{42} = 200 \text{ м.}$$

Из сопоставления результатов двух последних расчетов видно, что *следует проверять необходимую повторность опытов по каждому из факторов и брать наибольшую как предельную, обеспечивающую необходимую точность и надежность исследования.*

Изложенные примеры показывают особое значение комбинации считающих и записывающих приборов при опытах с выраженным влиянием случайных факторов. Вначале, используя считающие классификаторы, находят типичное для данных условий статистическое распределение измеряемых основных одной-двух переменных величин в течение относительно длительного промежутка времени работы классификатора. Если требуется особо высокая надежность и точность данных, из полученного распределения должно следовать, что значение его стандарта равно примерно одной шестой общей широты распределения (то есть разнице между максимальным и минимальным значением), что соответствует распределению, близкому к нормальному. В других случаях оценка получаемой при помощи считающих классификаторов выборки более сложна.

В условиях полевых опытов практически можно допустить, что считающие классификаторы должны работать в заданных условиях не менее одного часа, чтобы получить распределение, с высокой точностью и надежностью отображающее действительное развитие явлений. Затем по полученным распределениям судят, какие классы значений основных факторов в соответствии с целью исследования типичны для синхронной записи ряда изменяющихся величин в целях анализа функциональных связей. Зная время и путь работы считающих классификаторов, а также количество замеров по типичным классам, рассчитывают путь или время работы более сложных приборов для синхронной записи взаимосвязанных величин. Если показания классификатора снимать (или записывать) на определенных участках пути, то можно объективно выбрать участки для работы с записывающими приборами. Такая схема неизменно приводит к успеху: исследователь добивается высокой точности и надежности при наименьших затратах труда и средств.

Использование таблицы 1.13 многообразно и не ограничивается приведенными примерами.

Рассмотрим еще несколько важных случаев применения этой таблицы при распределении совокупности близко нормальному закону.

Иногда исследователю приходится решать задачи о всякого рода допусках, в том числе и по браковым отклонениям в качестве обработки детали или сельскохозяйственных материалов, причем необходимо опытным путем проверять качество.

Например, пусть требуется установить количество анализов семян для наладки технологического процесса сортировального пункта, если содержание мертвых примесей в семенах не должно превышать 3% (превышение приводит к рефакции), а методы измерения указанной засоренности дают значение стандарта ошибки в 0,2% от средней. Вместе с тем простое указание на необходимость полностью исключить рефакцию привело бы к переналадке технологического процесса или ко второму пропуску возможно вполне годной партии семян, что также невыгодно. Пусть из хозяйственных соображений (равенство убытков) допускается, что засоренность не должна превышать 2,8% с надежностью 0,95 и с ошибкой в 1,0 стандарта и в то же время не быть ниже 2,4 % с надежностью 0,80 и с ошибкой в 2,0 стандарта. Тогда для верхнего предела заданного интервала по таблице 1.13 следует взять 7 проб для анализа, получить среднюю, и, если эта средняя покажет засоренность 2,8% и ниже, партия считается пригодной для сдачи без переналадки машин или второго пропуска. Для нижнего предела требуется всего 2 пробы. Ясно, что в данном случае требуемую точность и надежность средних данных по заданным границам получают при наибольшем из вычисленных чисел и проб, то есть при 7 пробах.

Полученное в примере число анализов проб следует считать достаточным, но не необходимым. Так, например, если анализ первых двух проб из партии семян дал засоренность в 1,0% при разнице в результатах не более 0,2%, анализы вряд ли стоило бы продолжать. Определенный вывод можно было бы сделать также из 2—3 анализов, если по ним засоренность семян превышала бы с малой ошибкой, предположим 8%.

Таблицу 1.13 можно использовать также для приближенного установления *доверительных интервалов* (или доверительных границ), в которых с заданной надежностью (вероятностью) находится характеристика случайного процесса, в данном случае средняя. Иначе, в некоторых задачах ставится вопрос: если по нескольким опытам

вычислены средняя и стандарт, то какова вероятная ошибка этой средней в долях стандарта при заданной надежности?

Приведем краткий пример. Пусть в результате семи измерений средняя $\bar{a}=30$ мм, а стандарт $\sigma=0,5$ мм. Задаваясь надежностью (в данном случае возьмем $N=0,95$), находим ошибку средней долей стандарта. Ясно, что границы средней лежат в пределах $a\pm\sigma$. Как видно из таблицы 1.13, в этих условиях $\Delta=1$, то есть ошибка равна стандарту. Таким образом, с надежностью 0,95 можем принять, что близкое к истинному значению средней будет лежать в границах $30\pm 0,5$ мм.

Еще один пример. При контроле партии изготовленных деталей для измерений взято наудачу 11 деталей и по ним найдены: $\bar{a}=125$ мм и $\sigma=0,5$ мм. Спрашивается, с какой точностью и надежностью можно считать, что средний размер всей партии деталей лежит в пределах от 124,5 до 125,5 мм? Заметим, что доверительный интервал равен двум стандартам. Отсюда примем, что ошибки равна одному стандарту. Принимая это во внимание по таблице 1.13 находим, что при ошибке $\pm 0,5$ мм средний размер деталей всей партии будет находиться в предел $125\pm 0,5$ мм с надежностью 0,99.

Из таблицы 1.13 видим, что доверительные интервал растут с увеличением надежности и уменьшением числа опытов. Конечно, каждый исследователь стремится увеличить надежность результатов и уменьшить доверительные интервалы, что равносильно уменьшению ошибки вычисленной средней.

Есть два взаимосвязанных пути уменьшения доверительных интервалов.

1. Уменьшение численного значения стандарта, что при данном числе опытов (измерений) зависит только от применения более точных приборов и более совершенной методики опытов; так как более точные приборы дороже, то при этом увеличиваются прямые затраты на опыт за счет амортизации и ремонта этих приборов; если более точные приборы надежны — это лучший путь.

2. Увеличение числа опытов в тех же условиях, если это возможно; этот путь связан с дополнительными затратами времени и рабочей силы.

Подобные расчеты с использованием таблицы 1.13 в большинстве случаев требуют соответствующей интерполяции или после выбора значения надежности построения по таблице 1.13 плавной кривой количества опытов (в желательных пределах) в функции ошибки в долях стандарта.

О доверительных интервалах или границах подробно будет идти речь в разделе 2.

В связи с изложенным рассмотрим применение последовательного анализа.

*Последовательный анализ позволяет установить необходимую повторность измерений при заданных надежности и ошибке в долях стандарта, если сам стандарт неизвестен. Единственное обязательное условие при этом — **устанавливать по каждому измерению его результаты перед последующим измерением.***

Поэтому такой анализ вполне применим, если измерения проводятся при помощи инструментов или приборов, быстро дающих результаты (например, микрометров, манометров, считающих приборов с показом средней и стандарта и др.). Такой анализ весьма затруднителен при синхронных записях, требующих более или менее сложной расшифровки, например при осциллографировании. Упрощенное изложение последовательного анализа таково: после каждого измерения, начиная с третьего, вычисляют среднюю и стандарт. Как только достигается желательный результат (надежность и вычисленная ошибка в долях стандарта должны давать по таблице 1.13 то количество измерений, которое уже произведено), измерения следует прекратить. Приведем пример. Измерялось время, в течение которого определенный объем газа при постоянных давлении и температуре протекал через дроссель. К опыту были предъявлены требования: надежность 0,95 при ошибке в 0,5% от времени протекания газа. Из-за отсутствия данных было запланировано 10 повторностей. При первом измерении время истечения составило 75,4 сек, при втором—74,6 сек и при третьем— 75,0 сек. Средняя по трем измерениям составила 75,0 сек, а полупроцентная ошибка от средней— 0,375 сек. Вычисленный стандарт был равен 0,4 сек, то есть немногим больше шибки. В соответствии с таблицей 1.13 надо было бы провести больше семи измерений. Но следующее, четвертое, измерение дало время 75,1 сек. По четырем измерениям средняя составила 75,025 сек, полупроцентная ошибка средней почти не увеличилась, а значение стандарта упало до 0,358 сек. Таким образом, стандарт оказался меньше полупроцентной ошибки средней, что требовало по таблице 1.13 уже менее семи измерений. Пятое измерение дало значение времени также в 75,1 сек, опять очень малое увеличение ошибки и падение стандарта до 0,288. Шестое измерение дало время истечения в 74,9 сек и было последним, так как заданная ошибка стала больше стандарта примерно в 1,5 раза и заданная надежность была достигнута. В этих вычислениях была использована интерполяция 2.

Таким образом, упрощенный последовательный анализ с использованием таблицы 1.13 помог выполнить наложенные условия при экономии 40% запланированных измерений. В приведенном примере колебания результатов измерений были незначительны, и этим оправдывается упрощение последовательного анализа. Если колебания велики, следует пользоваться более сложным методом анализа. Наконец, следует указать, что применение последовательного анализа не обязательно влечет за собой сокращение предварительно намеченного числа измерений, а может привести и к их увеличению, если при предварительном расчете числа измерений были допущены ошибки.

Рассматривая таблицу 1.13, можно видеть, как быстро растет повторность опытов (измерений) при увеличении надежности и точности. Поэтому высокая точность измерений одной и той же величины дает возможность сильно уменьшить их количество при заданной надежности или повысить надежность при заданной повторности. Не всегда возможно увеличивать количество трудоемких или дорогих опытов. Если точность их известна, по таблице 1.13 устанавливают надежность результатов, а при подготовке и проведении опытов обеспечивают высокую точность измерений. В тех случаях, когда нет нужных данных и нельзя установить необходимую повторность, а затраты на поисковые опыты будут не меньше, чем на основные, часто принимают тройную повторность как минимальную.

Итак, **общая схема определения количества взаимосвязанных опытов такова.**

1. После того как установят количество основных факторов и подберут аппаратуру, определяют количество серий опытов m . Здесь под серией подразумевается группа опытов, проводимых при изменении значений одного или нескольких одновременно регистрируемых основных факторов. Количество серий m прямо пропорционально принятому общему количеству основных факторов $n_{\text{фо}}$ и обратно пропорционально количеству основных факторов $n_{\text{фс}}$, синхронно или одновременно регистрируемых:

$$m = \frac{n_{\text{фо}}}{n_{\text{фс}}} \quad (1.39)$$

Если при вычислениях m получается дробным, его округляют в большую сторону. Из формулы (1.39) следует, что уменьшить количество серий опытов можно двумя путями: уменьшая общее количество основных факторов $n_{\text{фо}}$ или увеличивая количество синхронно или одновременно регистрируемых факторов $n_{\text{фс}}$.

2. По каждой i -ой серии устанавливают число опытов n_{oi} , определяемое предполагаемыми формами наиболее сложных кривых функциональных зависимостей.

3. По всем основным факторам, значения которых будут измерены в какой-либо серии опытов, определяют их повторность p_i как максимальную при данной точности и заданной надежности.

В опытах для каждой отдельной серии повторность обычно одинакова. Тогда число опытов в i -ой серии при изменении одного из основных факторов равно:

$$N_{oi} = p_i n_{oi}.$$

Если же в данной серии опытов берут z градаций других основных факторов, то

$$N_{oi} = p_i \cdot n_{oi} \cdot z_i.$$

Во всех же m сериях исследования количество опытов будет:

$$N_0 = \sum_{i=1}^{i=m} p_i \cdot n_{oi} \cdot z_i. \quad (1.40)$$

Как правило, в каждом опыте данной серии количество измерений одно и то же, поэтому если в каждом опыте i -ой серии предположено произвести q_i измерений, то общее количество измерений в исследовании будет равно:

$$N_n = \sum_{i=1}^{i=m} p_i \cdot n_{oi} \cdot z_i \cdot q_i. \quad (1.41)$$

Из формулы (1.41) видно, что количество опытов и измерений растет в геометрической прогрессии, поэтому надо стремиться сузить исследование (в пределах цели), так как при методике, составленной небрежно и без учета особенностей однофакторных исследований можно необходимое количество измерений довести до десятков и даже сотен тысяч и сделать такое исследование неосуществимым. Разумеется, могут быть исследования, широту которых, то есть количество учитываемых факторов и измерений их величин, определений их свойств, нельзя уменьшить. Тогда необходимо принять меры к тому, чтобы возросли скорости измерений (не в ущерб их точности) и обработки опытных данных. Автоматизация и синхронизация измерений многих факторов, вычислительные машины для разработки планов опытов и обработки опытных данных дают возможность быстро проводить широкие исследования.

Вместе с тем, для многофакторных исследований существуют способы оценки значимости влияния от дельных факторов на развитие явления в процессе самих опытов, если есть возможность без значительных потерь времени получать результаты, или, как еще говорят, находить функцию отклика. Эту возможность используют для уменьшения количества опытов, отбрасывая второстепенные факторы, изменяя значение (уровень) меньшего количества факторов с иным шагом (интервалом) их изменения. Для уменьшения количества опытов в некоторых случаях можно использовать особенности дисперсионного или факториального анализа.

Из изложенного видно также практическое значение односторонне направленных исследований с выделением минимального количества основных факторов. Такого рода исследования могут иметь большое значение, когда надо определить функциональную зависимость явления от одного или немногих изменяющихся факторов, но могут и не дать общей закономерности развития явления, так как освещают лишь одну или немногие стороны его развития. Этот путь исследования можно было бы назвать аналитическим; он имеет самостоятельное теоретическое и практическое значение, но для познания явления в целом необходим синтез, надо установить закономерности, вытекающие из одновременного и совместного действия ряда факторов. В том случае, когда все опыты проводят по частным методикам, то при количестве r частных методик общее число опытов в исследовании в целом можно также представить как

$$N_0' + N_0'' + \dots + N_0 = \sum_{i=1}^{i=r} p_i n_{0i} z_i, \quad (1.42)$$

а общее количество измерений в исследовании

$$N_n = \sum_{i=1}^{i=r} p_i n_{0i} z_i q_i.$$

Продолжим рассмотрение первого примерера. Основным фактором является затупление лемеха. Выше мы приняли восемь опытов для описания кривой функции степени затупления лемеха (от нулевой до какой-то максимальной). Кроме этого, есть и другие основные факторы. Из них по разного рода соображениям отбираем еще два, подлежащих изменению: тип почвы и ее влажность. Условимся, что влияние всех этих факторов мы изучаем только для того, чтобы определить, можно ли резко изменить намеченные рабочей гипотезой закономерности, и поэтому достаточно двух градаций по каждому фактору (например, типы почв — чернозем и каштановая, влажность — нормальная спелость и 75% от нее). Таким образом, число градаций

других основных факторов $z=4$. Установленная выше поисковыми опытами длина участка для замеров (200 м) обеспечивает достаточные точность и надежность опытов, но неясно, какова будет надежность при изменении типа почвы и ее влажности, поэтому берем для них тройную повторность опытов.

Так как исследование однофакторное, то количество серий опытов $n=3$: $1=3$. В однофакторном исследовании в каждой серии изменяется только один фактор. Тогда по формуле (1.40) определим общее количество опытов по изложенной части методики первого примера:

$$N_{\text{оп}} = 3 \cdot 8 \cdot 4 = 96.$$

Все эти опыты, в целях нейтрализации всех остальных факторов, проводим на местности, приближающейся к горизонтальной, при ходах «туда», и «обратно», при одной и той же скорости агрегата, неизменных прицепе или навеске и постоянной установке плуга.

Кроме этого, чтобы более точно проследить, изменяются ли отклонения от средней в функции затупления лемехов при вспашке на одну и ту же глубину, назначим две серии специальных опытов по типам почв при нормальной спелости с таким изменением установки плуга, которое давало бы при всех градациях затупления лемехов примерно одну и ту же среднюю глубину пахоты. Значит, получим еще $8 \times 2 = 16$ опытов. Итого необходимо провести $96 + 16 = 112$ опытов.

Как следует из изложенного, в каждом опыте надо 420 раз измерить глубину вспашки, 2 раза при ходах («туда» и «обратно») — длину пути и время его прохождения; примем также 50 замеров ширины захвата плуга (это количество замеров не обосновываем, чтобы не загромождать изложение) и две динамограммы тягового сопротивления плуга. Если этим ограничиться и не принимать во внимание количество отметок о смене лемехов и качестве вспашки (оборот пласта, вспушенность и т. д.), в каждом опыте необходимо выполнить 474 измерения и две записи сопротивления плуга, а по всем опытам — 53 088 измерений и 224 динамограммы. Количество измерений можно значительно сократить, если ввести автоматические (не обязательно синхронные) измерения глубины вспашки и ширины захвата плуга. Отсюда видна значимость применения автоматически действующих измерительных приборов, в данном случае лучше считающими механизмами.

Несколько иначе следует подойти к определению количества опытов во втором примере. Не останавливаясь на подробных расчетах по результатам поисковых опытов, укажем, что для этого исследования было бы целесообразно провести в первой его стадии «чистые» лабораторные или лабораторно-полевые опыты на выровненной

горизонтальной площадке с искусственными (практически жесткими) препятствиями различной высоты и различной формы.

Основными факторами здесь являются: высота препятствия, его форма и скорость движения. Так как запись величин в этом исследовании автоматизирована и синхронизирована, то по существу проводятся серии опытов с одними и теми же способами измерений, но с изменением величин, характеризующих основные факторы. Если принять три значения высоты препятствия (нулевая, максимальная и средняя), три формы препятствия и движение на семи различных скоростях (все остальные величины неизменны), общее количество опытов в этой части исследования при одной повторности будет равно:

$$N_{\text{оп}} = 3 \cdot 3 \cdot 7 = 63 \text{ опыта.}$$

Если повторность больше единицы, число опытов соответственно увеличится. «Чистые» опыты дадут законы изменения кривых функций, полевые опыты — синтез уже выясненных закономерностей развития явлений. Выбираем условия полевых опытов:

- а) поле хорошо выровненное;
- б) поле с резко выраженным микрорельефом;
- в) движение на пяти передачах: две крайние и три в пределах допустимой (или оптимальной для данных условий) скорости;
- г) давление в шинах неизменное, близкое к оптимальному для данных условий (но может меняться по сериям опытов).

Таким образом, количество полевых опытов для второго примера будет равно:

$$N_{\text{оп}} = 2 \cdot 5 = 10.$$

Этот пример показывает, что при тщательно поставленных «чистых» опытах количество полевых, обычно более дорогих и трудоемких опытов в ряде случаев можно сократить.

Изложенный выше методический подход для определения количества опытов при проведении исследований практически может применяться во всех видах исследований независимо от их предметной направленности.

2. Погрешности измерений и устройств

2.1. Погрешности измерений

2.1.1. Измеряемые сигналы

Измеряемые сигналы (параметры, величины) являются случайными функциями времени. Они могут принимать различные значения, определяемые состоянием процесса, который ими характеризуется. На рис. 2.1 представлены отдельные реализации измеряемого сигнала в

функции времени, полученные при наблюдении за сигналом в течение интервала времени τ .

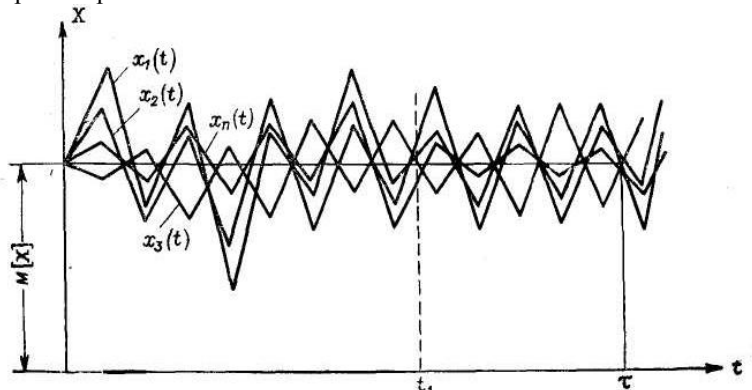


Рис. 2.1. Реализация случайного процессе.

Математическое ожидание $M[X]$ сигнала в данном случае величина постоянная. Непостоянство измеряемого сигнала в пределах одной реализации и отличие реализации друг от друга объясняются тем, что процесс, параметр (сигнал) которого измеряется, не может быть изолирован полностью от других процессов, совершающихся в том же пространстве. Так, высота горизонтального полета летательного аппарата будет случайным процессом, поскольку она определяется не только положением рулей высоты, но и турбулентностью атмосферы, действующей на летательный аппарат и изменяющейся случайным образом. Часто математическое ожидание сигнала не остается постоянным, а изменяется по какому-либо детерминированному закону (линейному, гармоническому или другому). При этом сигнал также будет случайной функцией и реализации его будут отличаться друг от друга за счет случайных изменений угла наклона (при линейном процессе), амплитуды и фазы (при гармоническом процессе), как это видно из рис. 2.2. В этом случае измеряемый сигнал является функцией времени, содержащей детерминированную и случайную составляющие. При этом детерминированная составляющая должна рассматриваться как математическое ожидание $M[X]$ процесса (сигнала).

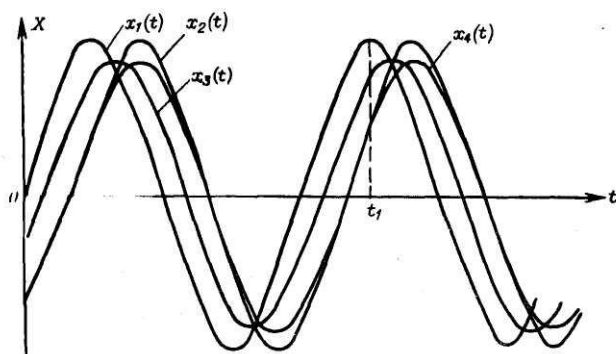


Рис. 2.2. Случайный процесс с математическим ожиданием, изменяющимся по гармоническому закону.

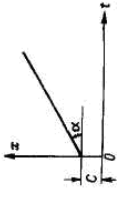
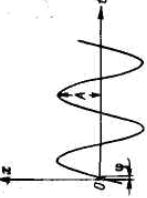
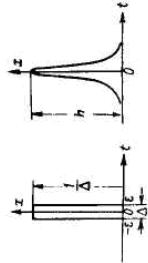
Каждая реализация в отдельности—это неслучайная функции, В отличие от случайного процесса $X(t)$ ее обозначают обычно (x_1) , (x_2) и т. д. Если произвести сечение случайного процесса в какой-либо момент времени t_1 (пунктирные линии на рис. 2.1 и 2.2) и рассматривать его только в этот момент, то случайный процесс превратится в случайную величину X , принимающую n различных значений (n — число реализаций). Таким образом, **случайная величина — это сечение случайной функции.**

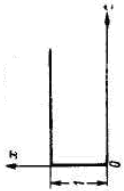
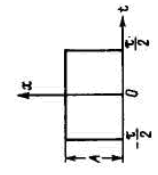
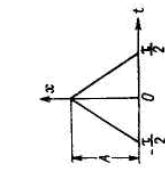
В табл. 2.1 приведены некоторые виды детерминированных составляющих сигналов (величин) в аналитической и графической форме, а в табл. 2.2— некоторые наиболее распространенные виды распределений случайных составляющих измеряемых сигналов (величин). Во второй графе табл. 2.2 даны аналитические выражения для плотности распределения вероятности появления отдельных значений сигнала (обычно говорят кратко — плотность распределения); а в третьей — графическая интерпретация плотностей распределения, дающая наглядное представление о вероятности появления измеряемой величины в различных интервалах значений. Необходимые пояснения даны в таблицах. Более подробно законы распределения и числовые характеристики случайных величин будут рассмотрены по отношению к случайным погрешностям измерений.

В первых двух разделах настоящего изложения исследуются погрешности измерения и погрешности измерительных устройств при условии, что измеряемый сигнал за время измерения остается неизменным. Погрешности измерительных устройств и измерений при этом являются случайными величинами, в общем случае обладающими отличными от нуля математическими ожиданиями (детерминированными составляющими).

Таблица 2.1

Некоторые виды детерминированных сигналов

Вид сигнала	Аналитическое выражение	Графическое изображение	Параметры	Модуль спектральной плотности. Фазовый спектр
Линейный	$x(t) = C + \text{tg } \alpha \cdot t \text{ при } t \geq 0$ $x(t) = 0 \text{ при } t < 0$		C, α	$S(\omega) = \frac{C}{\omega} - \text{tg } \alpha \frac{1}{\omega^2}$ $\psi(\omega) = \frac{\pi}{2}$
Синусоидальный	$x(t) = A \sin(\omega_0 t + \varphi) \text{ при } t \geq 0$ $x(t) = 0 \text{ при } t < 0$		A, ω_0 , φ	$S(\omega) = \frac{A\omega_0}{\omega_0^2 - \omega^2}$ $\psi(\omega) = 0$
Импульсная δ -функция	$\delta(t) = 0 \text{ при } t \neq 0; \delta(0) = \infty$ $S_0 = \int_{-\infty}^{\infty} \delta(t) dt = 1 \text{ при } z > 0$ <p>1) $\Delta \rightarrow 0$; 2) $\delta = \lim_{h \rightarrow \infty} \frac{h}{\sqrt{z}} e^{-ht^2}$</p> $\frac{1}{\Delta} \rightarrow \infty \text{ при } h \rightarrow \infty$		1) Δ 2) h	$S(\omega) = 1$ $\psi(\omega) = C$

<p>Единичная функция</p>	$x(t-\tau) = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases}$		<p>1</p>	$S(\omega) = \frac{1}{\omega}$ $\psi(\omega) = \frac{\pi}{2}$
<p>Прямоугольный импульс</p>	$x(t-\tau) = \begin{cases} A & -\frac{\tau}{2} \leq t \leq \frac{\tau}{2} \\ 0 & t < -\frac{\tau}{2}; t > \frac{\tau}{2} \end{cases}$		<p>A, τ</p>	$S(\omega) = \frac{2A}{\omega} \left \sin \frac{\omega\tau}{2} \right $ $\psi(\omega) = -\frac{\omega\tau}{2} \pm k\pi$
<p>Треугольный импульс</p>	$x(t) = \begin{cases} 2A\frac{t}{\tau} + A & -\frac{\tau}{2} \leq t \leq 0 \\ -2A\frac{t}{\tau} + A & 0 \leq t \leq \frac{\tau}{2} \\ 0 & t < -\frac{\tau}{2}; t > \frac{\tau}{2} \end{cases}$		<p>A, τ</p>	$S(\omega) = \frac{8A}{\omega^2\tau} \sin^2 \frac{\tau\omega}{4}$ $\psi(\omega) = 0$


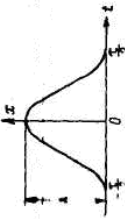
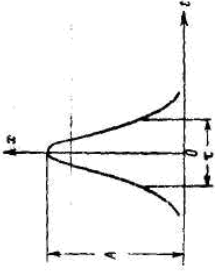
Вид сигнала	Аналитическое выражение	Графическое изображение	Параметры	Модуль спектральной плотности, фазовый спектр
Экспоненциальный импульс	$x(t) = \begin{cases} A \exp(-t/\tau) & t \geq 0 \\ 0 & t < 0 \end{cases}$		A, τ	$S(\omega) = \frac{A}{\sqrt{1 + \omega^2 \tau^2}}$ $\psi(\omega) = \text{arctg } \omega \tau$
Косинус-квадратный импульс	$x(t) = A \cos^2 \frac{\pi}{2} t$ <p style="text-align: center;">при $-\frac{\tau}{2} \leq t \leq \frac{\tau}{2}$</p>		A, τ	$S(\omega) = \frac{4\pi^2 A}{\omega} \left \frac{1}{4\pi^2 - \tau^2 \omega^2} \sin \frac{\omega \tau}{2} \right $ $\psi(\omega) = \frac{\omega \tau}{2} = k\tau$
Колоколообразный импульс	$x(t) = A \exp(-t^2/\tau^2)$ <p style="text-align: center;">$-\infty < t < \infty$</p>		A, τ	$S(\omega) = A \tau \sqrt{\pi} \exp(-\omega^2 \tau^2 / 4)$ $\psi(\omega) = 0$

Таблица 2.2

Плотности распределения случайных величин

Распределение	Аналитическое выражение	Графическое изображение	Параметры
Нормальное	$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp - \left[\frac{(x - m_x)^2}{2\sigma^2} \right]$ $-\infty < x < \infty$		m_x, σ
Равномерной плотности	$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha < x < \beta \\ 0 & x < \alpha; x > \beta \end{cases}$		m_x, σ
Арксинуса для сигнала $x = X_m \sin \omega t$ со случайной фазой	$f(x) = \frac{1}{\pi X_m \sqrt{1 - \left(\frac{x}{X_m}\right)^2}}$ $-X_m < x < X_m$...
Хи-квадрат	$f(x) = \begin{cases} 0 & x < 0 \\ \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) & x > 0 \end{cases}$ $\alpha > 0; \beta > 0$ $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} \exp(-y) dy$		α, β

2.1.2. Погрешности измерений

Измерение — это определение значения измеряемой величины (сигнала) опытным путем, путем эксперимента, поставленного в конечном счете для преобразования измеряемого сигнала в форму, удобную для наблюдения и использования.

Если через X_{00} обозначить истинное значение измеряемого сигнала, то при идеальном преобразовании результат измерения может быть представлен в виде

$$Y_{00} = k_n X_{00}, \tag{2.1}$$

где k_n - коэффициент преобразования, постоянный во всем диапазоне изменения измеряемой величины и принимаемый за номинальное значение чувствительности канала измерения.

В реальных условиях измерения выходной сигнал Y отличается от Y_{00} за счет непостоянства коэффициента преобразования, характеризуемого величиной Δk , и за счет внешних и внутренних помех ΔY_1 приведенных к выходу. Поэтому реальное уравнение преобразования следует записать как

$$\begin{aligned} Y &= (k_n + \Delta k) X_{00} + \Delta Y_1 = k_n X_{00} + \Delta k X_{00} + \Delta Y_1 = \\ &= k_n X_{00} + \Delta Y_2 + \Delta Y_1 = Y_{00} + \Delta Y, \end{aligned} \quad (2.2)$$

где $\Delta Y_2 = \Delta k X_{00}$ — ошибка, вызванная непостоянством коэффициента преобразования; ΔY — полная ошибка измерения, приведенная к выходу.

Тогда ошибка ΔX в определении измеряемой величины, или ошибка, приведенная ко входу, может быть определена из соотношения

$$X = \frac{1}{k_n} Y = X_{00} + \Delta X.$$

Итак, погрешность измерения

$$\Delta X = X - X_{00} \quad (2.3)$$

есть разность между полученным значением измеряемой величины и ее истинным значением. В дальнейшем X при Δ будем опускать, т. е. будем писать $\Delta = X - X_{00}$. Так как истинное значение измеряемой величины не известно, то и определить точно погрешность измерения нельзя. Можно лишь оценить погрешность измерения, если за X_{00} принять некоторое значение X_0 , настолько приближающееся к истинному, что для данной цели оно может быть использовано вместо него. Это значение X_0 измеряемой величины называется **действительным значением**. Тогда оценка погрешности измерения

$$\Delta = X - X_0 \quad (2.4)$$

(часто, чтобы отличить оценку от погрешности, ее обозначают $\hat{\Delta}$).

Погрешность измерения определяется *погрешностями используемой измерительной аппаратуры; погрешностью метода (при косвенных измерениях); погрешностями, вызванными несогласованием входных параметров измерительной аппаратуры и выходных параметров цепи измеряемого сигнала; погрешностями, вызванными несоответствием инерционных свойств измерительной аппаратуры и скорости изменения измеряемого сигнала (динамическая ошибка);*

погрешностями от неучтенного влияния внешних факторов на условия эксперимента.

Кроме того, оценка погрешности зависит от точности описания измеряемого сигнала $x(t)$, т. е. от степени приближения действительного значения к истинному (в случае статических измерений — приближения X_0 к X_{00}).

По характеру изменения погрешности измерения можно разделить на систематические и случайные.

Систематические погрешности — это погрешности постоянные или изменяющиеся по некоторому закону, который может быть достаточно просто установлен путем постановки соответствующего эксперимента.

При постановке эксперимента источники систематических ошибок должны быть по возможности исключены. Неисключенные систематические погрешности должны быть учтены путем введения в результат измерения поправок (исправленный результат измерения).

Случайные погрешности — это погрешности, изменяющиеся от опыта к опыту и не поддающиеся непосредственному учету вследствие их хаотического изменения, обусловленного одновременным воздействием на результат измерения ряда различных, не зависящих друг от друга факторов.

Наличие случайных погрешностей обнаруживается в том, что при многократном повторении измерения одной и той же величины в одинаковых условиях с одинаковой тщательностью результаты измерений несколько отличаются друг от друга. Случайные погрешности неизбежны. Поэтому следует считать недостаточно тщательными эксперименты, при повторении которых получаются совершенно одинаковые результаты. Следует заметить, что многие причины возникновения случайных ошибок нам известны (трение в опорах, неуравновешенность подвижных частей, порог чувствительности, дискретность и т. д.). Но от этого мало что меняется — погрешности, вызываемые ими, носят по-прежнему случайный характер.

При исследовании результатов наблюдений среди случайных погрешностей может встретиться промах.

Промах - это случайная погрешность, явно несоответствующая правильному результату измерения и свидетельствующая о грубой ошибке в постановке эксперимента, в отсчете.

После установления промаха он должен быть исключен из результатов измерений.

В общем случае *погрешность результата* имеет *систематическую и случайную составляющие*, даже если введены поправки на систематические погрешности, вызванные известными влияющими факторами. Объясняется это тем, что, во-первых, значения факторов не остаются в процессе измерения постоянными и, во-вторых, тем, что на результат измерения воздействуют факторы, действие которых в данном эксперименте не предполагалось, или факторы, действие которых невозможно учесть.

Так как в погрешность измерения входит случайная составляющая, то ее следует считать случайной величиной. На рис. 2.3 изображены реализация полной погрешности измерения в функции номера эксперимента или, в конечном счете, случайной функции времени.

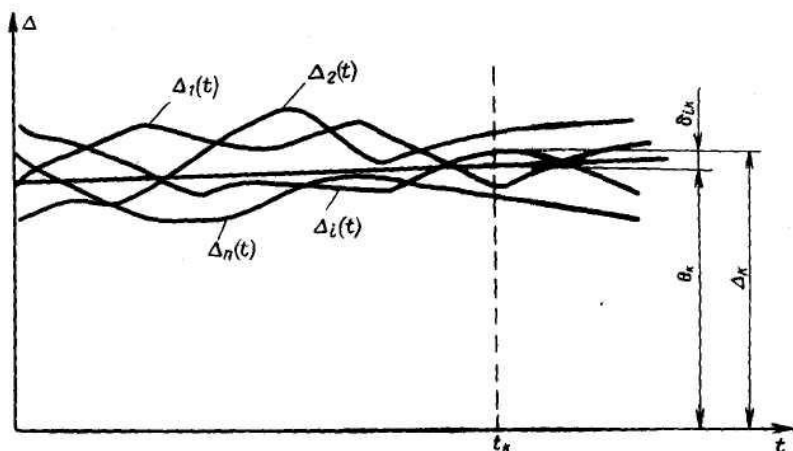


Рис. 2.3. К определению случайной погрешности.

Сечение функции есть, как уже говорилось, *случайная величина*. На рис. 2.3 сечение выполнено для момента времени t_k . Для каждого эксперимента (реализации) в сечении погрешность принимает разное, случайное значение Δ_{ik} , но все эти значения группируются около некоторого среднего значения θ_k . В других сечениях группирование погрешностей происходит около своих средних значений.

Очевидно, что указанные средние значения следует принять за систематические погрешности измерения в фиксированные моменты времени, а кривую, соединяющую их, следует считать зависимостью систематической погрешности от времени. Разность между погрешностью измерения и средним значением (систематической погрешностью) в каждом сечении следует рассматривать как

случайные погрешности $\delta_{1k}, \delta_{2k}, \dots, \delta_{ik}, \dots, \delta_{nk}$. Таким образом, случайная погрешность в момент измерения (например, t_k) может принять любое из значений Δ_{ik} . Значение полной погрешности измерения для любого момента времени будем записывать как

$$\Delta = \theta + \delta. \quad (2.5)$$

Систематическая составляющая погрешности при обработке результатов измерений должна быть определена, и на нее, как уже говорилось, должен быть исправлен результат измерения (исправленный результат измерений),

Точность измерения характеризуется случайной погрешностью. Определить ее можно аналитически и экспериментально.

При *аналитическом* определении проводится анализ источников погрешностей измерительной аппаратуры, метода измерения и др., определяется характер погрешностей, их величина, взаимное влияние. По полученным данным, используя аппарат суммирования частных погрешностей случайного характера и частных погрешностей систематического характера, можно найти погрешность измерения в нормальных условиях и сделать указания о дополнительных (систематических) погрешностях, которые могут возникнуть при отклонении условий измерения от нормальных.

Встает вопрос: можно ли утверждать, что рассчитанная на основании анализа погрешность измерения точно соответствует значению погрешности проведенного реального измерения? Конечно, нет. Путем анализа можно лишь оценить погрешность, так как нет возможности правильно учесть все факторы, определяющие погрешность измерения, и прежде всего законы распределения погрешностей и внешних факторов.

Экспериментальный метод определения погрешности измерения требует статистических данных, т. е. многократного повторения измерения в одинаковых условиях с одинаковой тщательностью. Путем обработки статистических данных можно найти случайную погрешность отдельного измерения, наиболее достоверное значение измеряемой величины (действительное значение) и погрешность его определения, т. е. установить границы, в которых лежит истинное значение измеряемой величины, конечно, с некоторой доверительной вероятностью. Статистический метод следует рассматривать не только как метод определения погрешности измерения, но и как метод повышения точности измерения. Не следует думать, что полученная оценка погрешности отдельного измерения является неизменной. При повторении того же эксперимента через некоторое время в тех же

условиях, с той же тщательностью можно получить и несколько другую оценку погрешности,

2.1.3. Случайные погрешности

Наиболее полно случайные погрешности могут быть описаны лишь с помощью закона распределения, в частности интегральной функцией распределения, являющейся одной из форм закона распределения $F(\Delta)$. *Случайные погрешности относятся к непрерывным случайным величинам, т. е. они могут принимать несчетное множество значений.* Перечислить все значения, записать их не представляется возможным. Поэтому, когда говорят о вероятности появления тех или иных погрешностей, имеют в виду вероятность P того, что погрешность либо будет меньше некоторого значения $P(\Delta < \Delta_1)$, либо лежит в некотором интервале значений $P(\Delta_1 < \Delta < \Delta_2)$. (В дальнейшем это всегда следует иметь в виду, даже если не будет сделано специальных оговорок). Здесь Δ — случайная погрешность, а $\Delta_1, \Delta_2, \Delta_i$ — значения, которые она может принимать (аналогично: X — случайная величина, а x — значения, которые она принимает).

Функция распределения погрешностей — это распределение вероятностей появления погрешностей разной величины в функции этой величины. Это можно записать как

$$F(\Delta_1) = P(\Delta < \Delta_1). \quad (2.6)$$

Для выяснения свойств функции распределения воспользуемся геометрической интерпретацией (рис. 2.4).

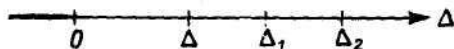


Рис. 2.4. Иллюстрация свойств функции распределения.

Будем рассматривать погрешность как точку Δ , занимающую на оси $O\Delta$ различные положения. Тогда $F(\Delta_1)$ есть вероятность того, что случайная погрешность Δ попадет левее Δ_1 [согласно (2.6)]. Если фиксирована точка $\Delta_2 > \Delta_1$, то $F(\Delta_2) \geq F(\Delta_1)$, т. е. функция распределения является неубывающей функцией своего аргумента. При перемещении Δ_1 вправо, в $+\infty$, попадание Δ левее Δ_1 становится достоверным событием, т. е. $F(+\infty) = 1$. Если перемещать точку Δ_1 влево, в $-\infty$, то вероятность того, что Δ попадет левее Δ_1 в пределе стремится к невозможному событию, т. е.

$$F(-\infty) = 0.$$

Зная вероятности появления Δ различных значений, можно построить график функции распределения (рис. 2.5,а).

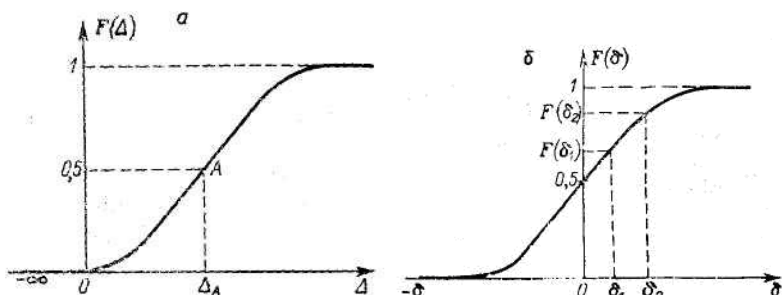


Рис. 2.5. Функция распределения: а—случайных погрешностей; б — случайной составляющей δ .

Если $F(\Delta)$ симметрична относительно точки A , соответствующей вероятности 0,5, то ее целесообразно сдвинуть по оси абсцисс на Δ_A , так как это подчеркивает симметричность наблюдаемых событий. Очевидно, такой сдвиг равносителен исключению систематической погрешности, равной $\theta = \Delta_A$. Поэтому функция распределения на рис. 2.5,б — это функция распределения только случайной составляющей погрешности. Вероятность попадания случайной погрешности в заданный интервал, например, $\delta_1 \leq \delta < \delta_2$ равна приращению функции распределения на этом интервале:

$$P(\delta_1 \leq \delta < \delta_2) = F(\delta_2) - F(\delta_1) \quad (2.7)$$

При сужении интервала, т. е. при $\delta_2 \rightarrow \delta_1$, вероятность $P(\delta = \delta_1 = \delta_2) \rightarrow 0$. Следовательно, для непрерывных случайных погрешностей вероятность появления погрешности любого определенного значения равна нулю. Это не означает, что такое событие невозможно; оно возможно, но вероятность его равна нулю,

Для описания случайных погрешностей можно воспользоваться также дифференциальной формой закона распределения—**плотностью распределения**

$$f(\Delta) = \frac{dF(\Delta)}{d\Delta}. \quad (2.8)$$

Плотность распределения, так же как и функция распределении, является одной из форм закона распределения, но **существует она только для непрерывных случайных величин**. Для анализа погрешностей эта форма закона распределения является более наглядной, поэтому широко используется для описания случайных погрешностей. График плотности распределения может иметь различную форму в зависимости от закона распределения

погрешностей. Для $F(\Delta)$, изображенной на рис. 2.5, $f(\Delta)$ имеет форму, близкую к форме колокола, как это изображено на рис. 2.6.

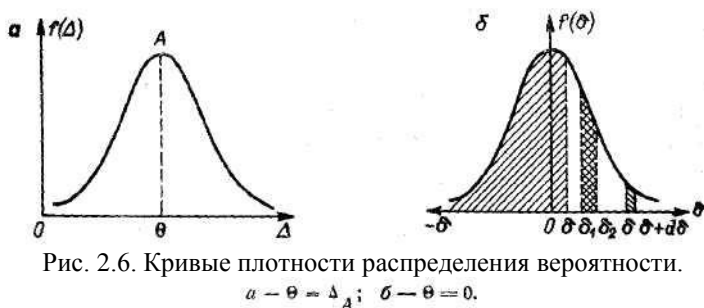


Рис. 2.6. Кривые плотности распределения вероятности.

$$a - \theta = \Delta_A; \quad \delta - \theta = 0.$$

Из рисунка хорошо видно, что в случае a погрешность имеет две составляющие: систематическую $\theta = \Delta_A$ и случайную δ , причем случайные погрешности группируются около θ ; малые случайные погрешности (относительно θ) имеют большую плотность, нежели большие. Кривая на рис. 2.6,б соответствует исправленным результатам измерения, когда на величину систематической ошибки внесена поправка. В этом случае систематическая погрешность равна нулю, и около нулевого математического ожидания группируются случайные погрешности.

Так как

$$F(\delta) = \int_{-\infty}^{\delta} f(\delta) d\delta, \quad (2.9)$$

то вероятность появления случайных погрешностей определяется площадью, ограниченной кривой $f(\delta)$ или ее частью и осью абсцисс, в зависимости от того, какой интервал погрешностей рассматривается. Так, (2.9) дает вероятность появления случайных погрешностей, лежащих в интервале от $-\infty$ до δ (здесь δ - некоторое предельное значение, которое может принимать случайная погрешность, но индекс опущен). Следовательно, на рис. 2.6,б это площадь под кривой, лежащая левее ординаты в точке δ (заштрихованная площадь).

Вероятность того, что погрешность лежит в интервале между δ_1 и δ_2 , равна площади, ограниченной ординатами в точках δ_1 и δ_2 и участками кривой $f(\delta)$ и оси абсцисс между ними. Аналитически это может быть выражено как

$$P(\delta_1 < \delta \leq \delta_2) = \int_{-\infty}^{\delta_2} f(\delta) d\delta - \int_{-\infty}^{\delta_1} f(\delta) d\delta = \int_{\delta_1}^{\delta_2} f(\delta) d\delta. \quad (2.10)$$

Величина $f(\delta)d\delta$ есть элемент вероятности, равный площади прямоугольника, опирающегося на $d\delta$. Так как

$$F(\delta) = \int_{-\infty}^{+\infty} f(\delta) d\delta = 1, \quad (2.11)$$

то площадь под всей кривой $f(\delta)$ равна единице, т. е. $F(\delta)$ есть **нормированная функция**. Плотность распределения — это неотрицательная функция, что непосредственно вытекает из того, что $F(\delta)$ есть неубывающая функция своего аргумента.

Выше уже говорилось, что закон распределения в любой форме является исчерпывающей вероятностной характеристикой случайных погрешностей как случайных величин, позволяющей оценить вероятность появления погрешностей разных значений, или плотность, с которой распределяются значения случайных погрешностей определенной величины. Однако не всегда удобно и необходимо характеризовать случайную погрешность таким исчерпывающим образом. Обычно указывают некоторые числовые характеристики случайных погрешностей, отражающие основные черты закона ее распределения. При анализе случайных погрешностей используются следующие основные числовые характеристики:

- 1) математическое ожидание, мода, медиана,
- 2) начальные и центральные моменты,
- 3) дисперсия и среднее квадратическое отклонение,
- 4) коэффициент асимметрии и эксцесс.

Большинство из названных числовых характеристик определяется через моменты.

2.1.4. Числовые характеристики

Математическое ожидание случайной погрешности — это среднее взвешенное ее значение, определяемое из значений Δ_i с учетом вероятностей их появления. Аналитически это можно выразить так:

- математическое ожидание

$$M[\Delta] = \frac{\Delta_1 P_1 + \Delta_2 P_2 + \dots + \Delta_i P_i + \dots + \Delta_n P_n}{P_1 + P_2 + \dots + P_i + \dots + P_n} =$$

$$= \frac{\sum_{i=1}^n \Delta_i P_i}{\sum_{i=1}^n P_i},$$

а так как

$$\sum_{i=1}^n P_i = 1,$$

то

$$M[\Delta] = \sum_{i=1}^n \Delta_i P_i. \quad (2.12)$$

Для погрешностей измерения как непрерывных величин выражение (2.12) можно записать в виде

$$m_{\Delta} = M[\Delta] = \int_{-\infty}^{+\infty} \Delta f(\Delta) d\Delta, \quad (2.13)$$

где $f(\Delta)d\Delta$ есть элемент вероятности.

Математическое ожидание погрешности измерения как случайной величины равно ее систематической погрешности.

Модой погрешности измерения называется ее наиболее вероятное значение, точнее, значение, плотность вероятности которого максимальна. На рис. 2.7 буквой М обозначена мода. В общем случае m и M не совпадают, Есть распределения одномодальные (рис. 2.7, а) и полимодальные (кривая распределения имеет несколько максимумов) (рис. 2.7,б), а также антимодальные, имеющие не максимум, а минимум (рис. 2.7,в).

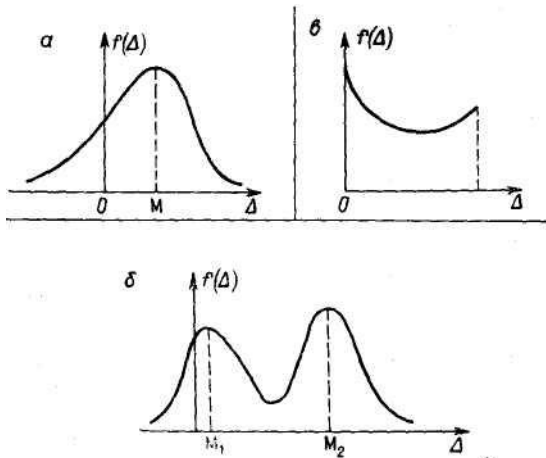


Рис. 2.7. Функции плотности законов распределения.
 а — одномодальный; б — двухмодальный; в — антимодальный.

Медианой Me погрешности измерения называется такое ее значение, для которого

$$P(\Delta < Me) = P(\Delta > Me),$$

т. е. одинаково вероятно, окажется ли случайная погрешность больше или меньше Me . Таким образом, медиана делит площадь под кривой плотности распределения пополам (рис. 2.8).

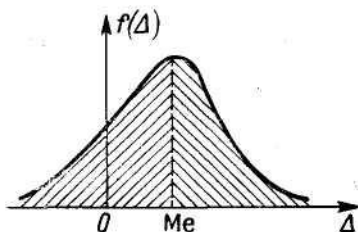


Рис. 2.8. Медиана случайной величины.

Начальным моментом s -го порядке непрерывной погрешности называется интеграл вида

$$\alpha_s [\Delta] = \int_{-\infty}^{+\infty} \Delta^s f(\Delta) d\Delta, \quad (2.14)$$

который может рассматриваться так же, как математическое ожидание s -й степени $M[\Delta^s]$, т. е. начальный момент можно представить в виде

$$\alpha_s [\Delta] = M [\Delta^s]. \quad (2.15)$$

В зависимости от номера s различают первый, второй и т. д. начальные моменты. Полагая $s = 1$, получим первый начальный момент

$$\alpha_1 [\Delta] = \int_{-\infty}^{\infty} \Delta f(\Delta) d\Delta, \quad (2.16)$$

равный, как это видно из (2.13), математическому ожиданию, т. е.

$$\alpha_1 [\Delta] = M[\Delta] = m_{\Delta} = \theta. \quad (2.17)$$

Разность между погрешностью измерения и ее математическим ожиданием есть случайная составляющая δ погрешности. Часто ее называют центрированной случайной погрешностью и обозначают

$$\delta = \Delta - m_{\Delta} = \Delta - \theta \quad (2.18)$$

Таким образом, случайная составляющая погрешности измерения— это **центрированная величина**. Моменты центрированных случайных величин называются **центральными моментами**. Центральным моментом s -го порядка погрешности измерения называется

математическое ожидание s -й степени центрированной погрешности измерения $\overset{\circ}{\Delta} = \delta$. Тогда центральный момент s -го порядка

$$\begin{aligned} \mu_s [\Delta] &= M [\overset{\circ}{\Delta}^s] = M [\delta^s] = M [(\Delta - m_\Delta)^s] = \\ &= \int_{-\infty}^{\infty} \delta^s f(\delta) d\delta. \end{aligned} \tag{2.19}$$

Очевидно, что первый центральный момент

$$\mu_1 [\Delta] = M [\overset{\circ}{\Delta}] = M [\delta] = 0. \tag{2.20}$$

Второй центральный момент

$$\mu_2 = M [\delta^2] = \int_{-\infty}^{\infty} \delta^2 f(\delta) d\delta. \tag{2.21}$$

Второй центральный момент называется также *дисперсией* этой величины — $D[\Delta]$. Дисперсия характеризует рассеивание погрешности измерения относительно ее математического ожидания. Она имеет размерность квадрата погрешности. Поэтому более наглядной характеристикой рассеивания является *среднее квадратическое отклонение*.

$$\sigma [\Delta] = \sigma_\Delta = \sqrt{D[\Delta]}. \tag{2.22}$$

Третий центральный момент μ_3 может рассматриваться как характеристика асимметрии распределения, так как для симметричного распределения все центральные моменты нечетного порядка равны нулю. Третий центральный момент имеет размерность куба, поэтому для получения безразмерной характеристики его делят на σ^3 . Тогда *коэффициент асимметрии*

$$Sk = \mu_3 / \sigma^3. \tag{2.23}$$

Четвертый центральный момент μ_4 используется для определения характеристики острровершинности (плосковершинности) распределения, называемой *эксцессом*:

$$Ex = \mu_4 / \sigma^4 - 3. \tag{2.24}$$

Так как вычитаемое в (2.24) является отношением μ_4 / σ^4 для нормального закона распределения, то эксцесс характеризует отклонение μ_4 / σ^4 для рассматриваемого закона от того же отношения для нормального и, следовательно, $Ex = 0$ для нормального распределения. На рис. 2.9 показаны соответственно плотности распределения с положительной и отрицательной асимметрией,

положительным и отрицательным эксцессом.

Рассмотренные числовые характеристики являются основными. Они чаще всего используются для приближенного описания случайных погрешностей (и случайных величин), при замене одного закона распределения другим с сохранением неизменными некоторых важнейших моментов.

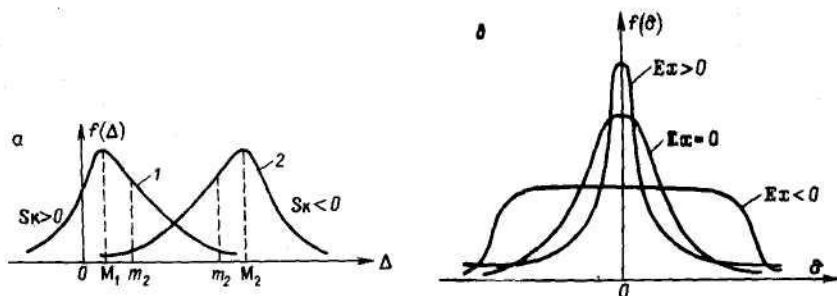


Рис. 2.9. Законы распределения.

a — с положительной и отрицательной асимметрией; *б* — с положительным и отрицательным эксцессом.

2.1.6. Нормальный закон распределения

Среди законов распределения нормальный закон занимает особое место, тем более для ошибок измерения. Дело в том, что погрешность измерения определяется большим числом частных составляющих, имеющих случайный характер, а из центральной предельной теоремы теории вероятностей следует, что распределение погрешностей измерения будет близко к нормальному, если результаты наблюдений формируются при воздействии большого числа независимо действующих частных погрешностей случайного характера, каждая из которых является незначительной по величине по сравнению с общей случайной погрешностью измерения (при рассмотрении законов распределения будем полагать $\theta=0$). Это имеет место и в том случае, когда частные погрешности имеют другие законы распределения, отличные от нормального. Чтобы закон распределения погрешностей измерения был близок к нормальному, число частных погрешностей должно быть, как показывает опыт, не менее 3—4.

Плотность распределения для нормального закона выражается как

$$f(\delta) = \frac{1}{\sigma \sqrt{2\pi}} \exp(-\delta^2/2\sigma^2) \tag{2.25}$$

и имеет вид колокола (рис. 2.10,*a*), симметричного относительно оси ординат, так как математическое ожидание $m=\theta$ (здесь в дальнейшем индексы при m , σ и других характеристиках будем опускать, если из текста понятно, что эти величины характеризуют). Нулю равны мода и медиана. Если математическое ожидание не равно нулю, то кривая будет сдвинута вправо или влево на величину m в зависимости от ее знака. Значение σ влияет на островершинность. Чем больше σ (см. рис. 2.10,*б*), тем меньше степень островершинности, тем вероятнее появление больших погрешностей и, наоборот, при уменьшении σ возрастает вероятность появления малых погрешностей, а больших — снижается.

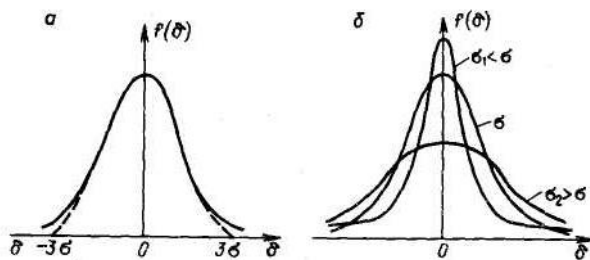


Рис. 2.10. Функция плотности нормального закона распределения (*a*) и изменение формы функции плотности в зависимости от значения σ (*б*).

Общее выражение для центральных моментов нормального распределения можно записать в виде

$$\mu_s = \int_{-\infty}^{\infty} \delta^s f(\delta) d\delta = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} \delta^s \exp(-\delta^2/2\sigma^2) d\delta. \quad (2.26)$$

Интеграл от (2.25) в элементарных функциях не выражается и может быть вычислен лишь с помощью таблиц; моменты μ_s можно найти достаточно легко, так как существует простое рекуррентное соотношение, позволяющее выражать моменты высших порядков через моменты низших порядков :

$$\mu_s = (s - 1) \sigma^2 \mu_{s-2}. \quad (2.27)$$

Введя понятие нулевого момента μ_0 , равного единице, как математическое ожидание нулевой степени, получим:

$$\mu_0 = 1; \quad \mu_1 = 0; \quad \mu_2 = \sigma^2; \quad \mu_3 = 0; \quad \mu_4 = 3\sigma^4. \quad (2.28)$$

Коэффициент асимметрии для нормального закона

$$Sk = \mu_3/\sigma^3 = 0. \quad (2.29)$$

Экцесс

$$Ex = (\mu_4/\sigma^4) - 3 = 0. \quad (2.30)$$

Для определения вероятности появления погрешностей δ , распределенных по нормальному закону, в заданном интервале функцию распределения

$$F(\delta) = \int_{-\infty}^{\delta} f(\delta) d\delta = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\delta} \exp(-\delta^2/2\sigma^2) d\delta$$

преобразуем, выразив δ в долях σ : $\delta = z\sigma$, или $z = \delta/\sigma$. Тогда функция распределения для нормально распределенной погрешности

$$\Phi^*(\delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\delta/\sigma} \exp(-z^2/2) dz. \quad (2.31)$$

Для нее составлены таблицы, по которым и определяется вероятность появления погрешностей в интервале от $-\infty$ до δ .

При определении вероятностей появления случайных погрешностей, лежащих в некотором интервале, удобнее пользоваться интегралом вероятности в виде

$$\Phi_1(\delta) = \Phi^*(\delta) - 0,5 = \frac{1}{\sqrt{2\pi}} \int_0^{\delta/\sigma} \exp(-z^2/2) dz. \quad (2.32)$$

Вероятность появления погрешностей, меньших $\pm\delta$, будет равна $2\Phi_1(\delta)$. Вероятность появления погрешностей в интервале $\delta_2 - \delta_1$

$$P[\delta_1 < \delta \leq \delta_2] = \Phi_1(\delta_2) - \Phi_1(\delta_1). \quad (2.33)$$

Наряду с функцией нормального распределения для выражения точности измерения часто используют стандартную аппроксимацию функции нормального распределения, полученную из условия, что $P=1$ соответствует $\delta_{\max} = \pm 3\sigma$ (см. пунктирную линию на рис. 2.13).

2.1.6. Закон равномерной плотности распределения

Выше уже говорилось, что общая погрешность измерения, как результирующая, чаще всего имеет нормальный закон распределения. Однако, если составляющие погрешности измерения имеют другие законы распределения и вес этих погрешностей в общей погрешности измерения неодинаков, закон распределения ее может оказаться и отличным от нормального, даже если число составляющих велико. Наиболее часто отдельные погрешности измерения (так же как и частные составляющие погрешностей измерительных устройств)

оказываются распределенными по нормальному закону, закону равномерной плотности, закону арксинуса со случайной фазой и другим, причем наиболее часто, кроме нормального, в измерениях приходится иметь дело с законом равномерной плотности.

Закон равномерной плотности распределения характерен для таких погрешностей измерения, как погрешность отсчета, погрешность дискретности, порог чувствительности, т. е. таких, которые могут принимать с равной плотностью вероятности любые значения, лежащие в некотором заданном интервале. Из рис. 2.11,а видно, что плотность распределения $f(\delta)$ в интервале от $+a$ до $-b$ остается постоянной, равной c , вероятность же появления малых погрешностей меньше, чем больших.

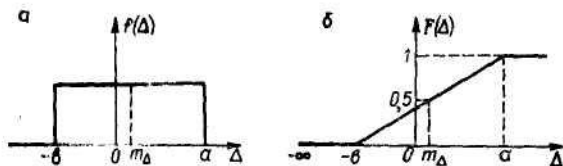


Рис 2.11. Закон равномерной плотности распределения (а) и его функции (б).

Для плотности распределения можно записать (для $m_{\Delta} = 0$):

$$f(\delta) = \begin{cases} c & \text{при } -b \leq \delta \leq a, \\ 0 & \text{при } \delta < -b; \delta > a, \end{cases} \quad (2.34)$$

или, так как площадь $c(a + b) = 1$, то $c = 1/(a + b)$ и

$$f(\delta) = \begin{cases} \frac{1}{a + b} & \text{при } -b \leq \delta \leq a, \\ 0 & \text{при } \delta < -b; \delta > a. \end{cases} \quad (2.35)$$

Тогда функция распределения

$$F(\delta) = \int_{-\infty}^{\delta} f(\delta) d\delta = \begin{cases} 1 & \text{при } \delta < -b, \\ f(\delta) [\delta - (-b)] = \frac{\delta + b}{a + b} & \text{при } -b \leq \delta < a, \\ 0 & \text{при } \delta \geq a. \end{cases} \quad (2.36)$$

График функции распределения приведен на рис. 2.11,б. Если $b = a$, то $f(\delta) = 1/2a$ и $F(\delta) = (\delta + a)/2a$ в пределах интервала $\pm a$.

Числовые характеристики случайной погрешности, подчиненной закону равномерной плотности в интервале от $+a$ до $-b$ находятся достаточно просто.

Математическое ожидание

$$m_{\Delta} = \int_{-b}^{+a} \frac{1}{a+b} \delta d\delta = \frac{a-b}{2} \quad (2.37)$$

(при $b = a$ $m_{\Delta}=0$; $\Delta = \delta$).

Медиана совпадает с m_{Δ} ; моды закон равномерной плотности не имеет. Второй момент, или дисперсия

$$\begin{aligned} \mu_2 = D_{\Delta} &= \int_{-b}^a \frac{1}{a+b} (\Delta - m_{\Delta})^2 d\Delta = \frac{1}{a+b} \int_{-b}^a \left(\Delta - \frac{a-b}{2} \right)^2 d\Delta = \\ &= \frac{(a+b)^2}{12} \end{aligned} \quad (2.38)$$

(при $b = a$ $D=a^2/3$).

Следовательно, среднеквадратическое отклонение

$$\sigma = \frac{a+b}{2\sqrt{3}} \quad (2.39)$$

(при $b = a$ $\sigma = a/\sqrt{3}$).

Четвертый момент

$$\mu_4 = \frac{1}{a+b} \int_{-b}^a \left(\Delta - \frac{a-b}{2} \right)^4 d\Delta = \frac{(a+b)^4}{80}. \quad (2.40)$$

Экцесс

$$Ex = \frac{(a+b)^4 \cdot 144}{(a+b)^4 \cdot 80} - 3 = -1,2. \quad (2.41)$$

Коэффициент асимметрии равен нулю.

Максимальная погрешность как максимальное отклонение от математического ожидания

$$\delta_{\max} = \pm \frac{a+b}{2}. \quad (2.42)$$

Так как $\sigma = (a+b)/2\sqrt{3}$, то $a+b = 2\sqrt{3}\sigma$ и

$$\delta_{\max} = \pm \sqrt{3} \sigma. \quad (2.43)$$

(Для нормально распределенных погрешностей обычно принимают $\delta_{\max} = \pm 3\sigma$; $P = 0,997$.)

Вероятность появления ошибок, лежащих в интервале $\pm d$ для $d < a$ определяется как

$$P[-d < \delta < d] = \int_{-d}^d \frac{1}{a+b} d\delta = \frac{2d}{a+b}, \quad (2.44)$$

т.е. путем умножения плотности распределения на интервал.

2.1.7. Композиции законов распределения. Правила суммирования случайных погрешностей

Имеется в виду вопрос о распределении плотности вероятности суммы нескольких независимых случайных величин, что очень существенно для определения закона распределения общей погрешности измерения, когда она складывается из небольшого числа составляющих. Допустим, что погрешность измерения $\Delta = \Delta_1 + \Delta_2$, причем составляющие имеют плотности распределения соответственно $f_1(\Delta_1)$ и $f_2(\Delta_2)$. Совместная плотность распределения двух независимых случайных погрешностей

$$f(\Delta_1, \Delta_2) = f_1(\Delta_1) \cdot f_2(\Delta_2), \quad (2.45)$$

и вероятность появления Δ , функция ее распределения,

$$\begin{aligned} F(\Delta) &= \iint_S f_1(\Delta_1) \cdot f_2(\Delta_2) d\Delta_1 d\Delta_2 = \\ &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\Delta - \Delta_2} f_1(\Delta_1) f_2(\Delta_2) d\Delta_2 \right\} d\Delta_1. \end{aligned} \quad (2.46)$$

Рис. 2.12 дает наглядную интерпретацию $F(\Delta)$ как площади S , лежащей левее прямой $\Delta = \Delta_1 + \Delta_2$ (здесь случайные погрешности и значения, которые они могут принимать, обозначены одинаково).

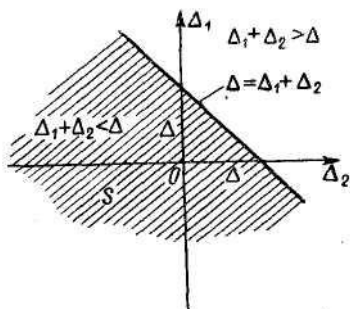


Рис. 2.12. К определению функции распределения двух погрешностей.

Область правее прямой соответствует нормальному случаю, когда сумма погрешностей больше общей погрешности; область левее прямой — случаю $\Delta_1 + \Delta_2 < \Delta$. Дифференцируя (2.46) по Δ , входящей в верхний предел внутреннего интеграла, получим плотность распределения суммы двух независимых случайных величин:

$$f(\Delta) = \int_{-\infty}^{\infty} f(\Delta_1, \Delta - \Delta_1) d\Delta_1,$$

$$f(\Delta) = \int_{-\infty}^{\infty} f(\Delta_2, \Delta - \Delta_2) d\Delta_2. \quad (2.47)$$

Выражения (2.47) являются равнозначными.

Найдем плотность распределения погрешности, являющейся суммой двух независимых случайных погрешностей Δ_1 и Δ_2 , с нормальными законами распределения:

$$f_1(\Delta_1) = \frac{1}{\sigma_1 \sqrt{2\pi}} \times \exp \left[-\frac{(\Delta_1 - m_{\Delta_1})^2}{2\sigma_1^2} \right],$$

$$f_2(\Delta_2) = \frac{1}{\sigma_2 \sqrt{2\pi}} \times \exp \left[-\frac{(\Delta_2 - m_{\Delta_2})^2}{2\sigma_2^2} \right], \quad (2.48)$$

где m_{Δ_1} , m_{Δ_2} , σ_1 , σ_2 — математические ожидания и среднее квадратические отклонения погрешностей Δ_1 и Δ_2 .

Согласно (2.47) можно написать плотность распределения суммы Δ_1 и Δ_2 :

$$f(\Delta) = \int_{-\infty}^{\infty} f_1(\Delta_1) f_2(\Delta - \Delta_1) d\Delta_1 =$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp \left[-\frac{(\Delta_1 - m_{\Delta_1})^2}{2\sigma_1^2} - \frac{(\Delta - \Delta_1 - m_{\Delta_2})^2}{2\sigma_2^2} \right] d\Delta_1.$$

После интегрирования и преобразований получим

$$f(\Delta) = \frac{1}{\sqrt{2\pi} \sqrt{\sigma_1^2 + \sigma_2^2}} \exp \left[\frac{[\Delta - (m_{\Delta_1} + m_{\Delta_2})]^2}{2(\sigma_1^2 + \sigma_2^2)} \right]$$

$$= \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{(\Delta - m_{\Delta})^2}{2\sigma^2} \right]. \quad (2.49)$$

Таким образом, закон распределения суммы двух независимых случайных величин, распределенных по нормальному закону, также является нормальным с математическим ожиданием

$$m_{\Delta} = m_{\Delta_1} + m_{\Delta_2}. \quad (2.50)$$

и средним квадратическим отклонением

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}. \quad (2.51)$$

Полученный результат может быть распространен и на случай суммы n случайных независимых погрешностей, распределенных по нормальному закону, что позволяет сформулировать правила суммирования случайных погрешностей измерения. Если составляющие погрешности измерения заданы или определены как $\Delta_1, \Delta_2, \Delta_3, \dots, \Delta_n$ с вероятностями $P_1, P_2, P_3, \dots, P_n$ и являются независимыми случайными величинами с нулевыми математическими ожиданиями, то для определения суммарной погрешности измерения $\Delta = \delta$ нужно поступить следующим образом.

1. По заданным составляющим погрешностям и вероятностям их появления найти средние квадратические отклонения. Для этого по приложению 1 для заданной вероятности P_i появления погрешности δ_i находят коэффициент z_i . Тогда

$$\sigma_i = \delta_i / z_i. \quad (2.52)$$

2. Определить среднее квадратическое отклонение полной погрешности измерения как суммы n независимых случайных величин, т. е.

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_i^2 + \dots + \sigma_n^2}. \quad (2.53)$$

3. По заданной вероятности P определения погрешности измерения найти по приложению 1 z . Тогда

$$\delta_i = z\sigma. \quad (2.54)$$

Композиция двух законов равномерной плотности дает так называемый закон Симпсона. Пусть имеются две независимые случайные погрешности — δ_1 и δ_2 , плотности распределения которых соответственно равны

$$f_1(\delta_1) = \frac{1}{2a} \text{ и } f_2(\delta_2) = \frac{1}{2b}$$

(рис. 2.13).

Плотность распределения

$$f(\delta) = \begin{cases} \frac{1}{2a} & \text{при } |\delta| < a - b, \\ \frac{a + b - |\delta|}{4ab} & \text{при } a - b \leq |\delta| \leq a + b, \\ 0 & \text{при } |\delta| > a + b. \end{cases} \quad (2.55)$$

По форме $f(\delta)$ -трапеция. При $b = a$ (см. рис. 2.13, б).

$$f(\delta) = \begin{cases} \frac{2a - |\delta|}{4a^2} & \text{при } |\delta| \leq 2a, \\ 0 & \text{при } |\delta| > 2a. \end{cases} \quad (2.56)$$

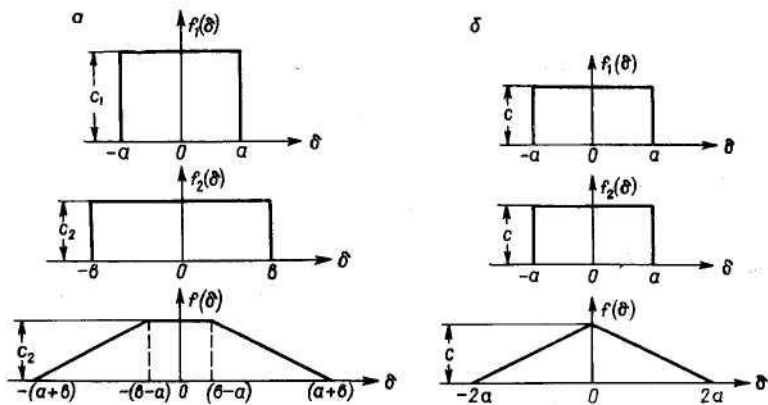


Рис. 2.13. Композиции законов равномерной плотности.

Среднее квадратическое отклонение

$$\sigma = \sqrt{\sigma_1^2 + \delta_2^2}. \quad (2.57)$$

Погрешность измерения

$$\delta = z\sigma,$$

где z —коэффициент, определяющий доверительный интервал для заданной вероятности.

Таким образом, закон равномерной плотности является неустойчивым. Композиция таких законов дает другой закон распределения, более близкий к усеченному закону нормального распределения, чем к закону равномерной плотности. При $n > 2$ приближение к нормальному закону становится все заметнее. Поэтому часто при числе частных погрешностей измерения $n \geq 3$, распределенных по законам равномерной плотности, для суммарной погрешности принимается закон нормального распределения.

Значительно сложнее определить закон распределению суммарной погрешности, если отдельные составляющий распределены по разным законам.

Рассмотрим композицию нормального закона и закона равномерной плотности. Пусть δ_1 распределена по нормальному закону, а δ_2 — по закону равномерной плотности. Тогда

$$f_1(\delta_1) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left(-\frac{\delta_1^2}{2\sigma_1^2}\right), \quad m_1 = 0,$$

$$f_2(\delta_2) = \frac{1}{2a} \quad \text{при } -a \leq \delta \leq a. \quad (2.58)$$

Используя (2.47), найдем для $\delta = \delta_1 + \delta_2$

$$f(\delta) = \frac{1}{2a} \int_{-a}^a \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left[-\frac{(\delta - \delta_2)^2}{2\sigma_1^2}\right] d\delta_2 =$$

$$= \frac{1}{2a} \int_{-a}^a \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left[-\frac{(\delta_2 - \delta)^2}{2\sigma_1^2}\right] d\delta_2 =$$

$$= \frac{1}{2a} \left[\frac{1}{\sigma_1 \sqrt{2\pi}} \int_{-\infty}^{-a} \exp\left[-\frac{(\delta_2 - \delta)^2}{2\sigma_1^2}\right] d\delta_2 - \right.$$

$$\left. - \frac{1}{\sigma_1 \sqrt{2\pi}} \int_a^{\infty} \exp\left[-\frac{(\delta_2 - \delta)^2}{2\sigma_1^2}\right] d\delta_2 \right]. \quad (2.59)$$

Композиция нормального и равномерной плотности распределений дает, таким образом, закон, отличный от нормального. На рис. 2.14 изображены плотности распределений $f_1(\delta_1)$, $f_2(\delta_2)$ и плотность распределения суммарной погрешности $f(\delta)$.

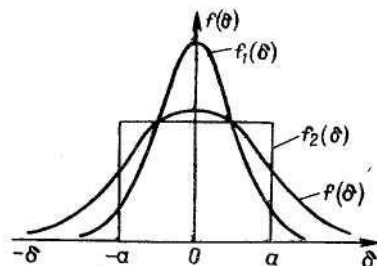


Рис. 2.14. Композиция законов равномерной плотности и нормального.

Суммарное среднее квадратическое отклонение $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$,
суммарная погрешность измерения $\delta = z\sigma$.

Значение z определяется законом распределения суммарной погрешности и заданной доверительной вероятностью. Можно показать, что при изменении «веса» погрешности δ_2 в общей погрешности измерения δ коэффициент z может принимать различные значения, например для $P = 0,997$ — от 3 до 1,73. Характеристикой «веса» может служить отношение средних квадратических отклонений рассматриваемой i -й погрешности и суммарной, т. е.

$$A_i = \sigma_i / \sigma. \quad (2.60)$$

Для композиции большого числа отдельных погрешностей, распределенных по известным и неизвестным, но симметричным законам, имеющим приблизительно одинаковый вес, может быть принят нормальный закон распределения. Тогда z выбирается по заданной вероятности, которая в свою очередь задается в зависимости от ответственности результатов измерений.

Если же среди отдельных составляющих погрешностей есть так называемые доминирующие погрешности, имеющие больший вес по сравнению с другими, и закон распределения их отличается от нормального, то вопрос о выборе может быть решен только на основании экспериментально полученных статистических данных, достаточных для построения закона распределения.

2.1.8. Зависимые и независимые случайные величины.

Корреляция случайных погрешностей

Уточним, что понимается в теории вероятностей под зависимыми и независимыми случайными величинами, в том числе и случайными погрешностями. Случайные величины X и Y называются независимыми, если закон распределения каждой из них не зависит от того, какое значение приняла другая случайная величина. Если это условие не выполняется, то случайные величины являются зависимыми, причем зависимыми в вероятностном смысле. Это значит, что случайные величины X и Y являются зависимыми, если значению одной величины X соответствует как-то значение другой величины Y , но как — неизвестно, т. е. нельзя сказать, какие значения будут принимать Y при изменениях величины X . Можно лишь сказать, что X и Y связаны вероятностной (стохастической) зависимостью или что закон распределения Y зависит от того, какое значение приняла величина X . Так, погрешности приборов, входящих в схему измерения, обычно принимаемые за независимые случайные

величины, при колебаниях температуры T окружающей среды оказываются связанными между собой вероятностной зависимостью, если температурное поле в месте расположения приборов нестационарно. Если температурное поле можно принять стационарным, то результат измерения будет содержать систематическую погрешность, являющуюся вполне определенной функцией температуры в точке ее измерения.

Таким образом, понятие «зависимость» в теории вероятностей отлично от понятия «функциональная зависимость». Оно является более широким понятием, справедливым для любой степени жесткости этой зависимости. Очевидно, что максимально предельная степень жесткости соответствует функциональной зависимости, минимальная — независимости.

Для системы независимых случайных величин плотность распределения системы равна произведению плотностей распределения величин, входящих в систему, т. е.

$$f(x,y)=f_1(x)\cdot f_2(y) \quad (2.61)$$

Это условие является необходимым и достаточным для установления независимости случайных величин. Если имеется система, характеризующаяся $f(x, y)$, то величины, входящие в нее, можно считать независимыми, если удастся путем математических преобразований показать, что $f(x, y) = f_1(x) \cdot f_2(y, x)$. Чаще вопрос о том, являются случайные величины зависимыми или нет, приходится решать на основании изучения самих величин и результатов их одновременного воздействия на тот или иной объект изучения, как это было сделано в рассмотренном примере. Случайная погрешность Δ_T измерения, вызванная изменением температуры, складывается из ряда составляющих: Δ_{TU} (источник питания), Δ_{TR} (сопротивление), Δ_{TN} (нуль — орган) и т. д. Эти составляющие погрешности связаны между собой через T и, следовательно, принципиально являются зависимыми.

В практике измерений одинаково часто приходится иметь дело с зависимыми и независимыми погрешностями. Так, погрешности Δ_{kX} и Δ_{kY} , вызванные нестабильностью коэффициентов усиления каналов горизонтального и вертикального отклонения электронного осциллографа, являются зависимыми, так как нестабильность их вызывается не только нестабильностью элементов усилителей, но и нестабильностью Δ_U напряжения источника питания. Погрешность же дискретности, погрешность отсчета и погрешность, вызванная нестабильностью чувствительности, — величины независимые. Вообще вероятностная зависимость составляющих погрешностей измерения

между собой чаще всего возникает через какой-либо влияющий фактор. Частным случаем вероятностной зависимости является линейная вероятностная зависимость, характеризующаяся тем, что при изменении одной случайной величины другая имеет тенденцию возрастать или убывать по вероятностному линейному закону. Эта тенденция к линейной зависимости может быть более или менее ярко выраженной, так же как всякая вероятностная зависимость. Максимальной предельной вероятностной линейной зависимостью, очевидно, является функциональная линейная зависимость. Степень тесноты (жесткости) линейной зависимости между случайными величинами характеризуется коэффициентом корреляции

$$r_{X,Y} = \frac{K_{X,Y}}{\sigma_X \sigma_Y}, \quad (2.62)$$

где $K_{X,Y}$ — корреляционный момент случайных величин X и Y (момент связи).

Корреляционный момент — это второй смешанный центральный момент:

$$\mu_{X,Y} = M[\hat{X}, \hat{Y}] = M[(X - m_X)(Y - m_Y)]. \quad (2.63)$$

Для непрерывных случайных величин X и Y

$$K_{X,Y} = \iint_{-\infty}^{\infty} (x - m_X)(y - m_Y) f(x,y) dx dy. \quad (2.64)$$

Для независимых случайных величин корреляционный момент равен нулю, что легко получить, если (2.64) переписать в виде

$$K_{X,Y} = \int_{-\infty}^{\infty} (x - m_X) f_1(x) dx \int_{-\infty}^{\infty} (y - m_Y) f_2(y) dy. \quad (2.65)$$

Здесь каждый интеграл является первым центральным моментом (величин X и Y) и, следовательно, равен нулю.

Обратное положение — если корреляционный момент равен нулю, то случайные величины независимы — является несправедливым. Равенство корреляционного момента нулю является необходимым, но недостаточным условием независимости случайных величин. Можно найти примеры, когда при $K_{XY} = 0$ величины X и Y являются зависимыми.

Коэффициент корреляции, являющийся характеристикой степени тесноты линейной связи между случайными величинами, может принимать значения в пределах

$$-1 \leq r_{X,Y} \leq 1. \tag{2.66}$$

Предельные значения соответствуют положительной и отрицательной линейной функциональной зависимости.

При определении погрешности измерения, складывающейся из зависимых случайных составляющих, предполагается всегда линейная вероятностная зависимость. Это дает возможность учесть степень этой зависимости при суммировании составляющих путем введения коэффициента корреляции. Так, если случайные составляющие погрешностей δ_1 и δ_2 характеризуются средними квадратическими отклонениями σ_1 и σ_2 , а коэффициент корреляции δ_1 и δ_2 равен $r_{1,2}$, то среднее квадратическое отклонение суммы случайных погрешностей

$$\sigma = \sqrt{\sigma_1^2 + 2r_{1,2}\sigma_1\sigma_2 + \sigma_2^2}. \tag{2.67}$$

Это легко доказать, если вспомнить, что квадрат среднего квадратического отклонения есть дисперсия и что дисперсия суммы двух случайных величин $X + Y = Z$

$$\begin{aligned} D[X + Y] = D[Z] &= M[\dot{Z}^2] = M[\dot{X}^2] + 2M[\dot{X}\dot{Y}] + M[\dot{Y}^2] = \\ &= D[X] + 2K_{X,Y} + D[Y], \end{aligned} \tag{2.68}$$

или, так как

$$K_{X,Y} = r_{X,Y}\sigma_X\sigma_Y, \tag{2.69}$$

то

$$D[Z] = \sigma_z^2 = \sigma_x^2 + 2r_{X,Y}\sigma_X\sigma_Y + \sigma_Y^2. \tag{2.70}$$

В дальнейшем (2.70) будем называть законом суммирования двух зависимых (точнее, коррелированных) случайных величин. Определение коэффициента корреляции возможно, если известна совместная плотность распределения $f(x, y)$, $f(\Delta_1, \Delta_2)$. Если $f(\Delta_1, \Delta_2)$ не известна и определение ее затруднительно, то на основании анализа тесноты связи между Δ_1 и Δ_2 принимают коэффициент корреляции равным ± 1 или 0. При $r_{1,2}=0$ (2.67) принимает вид закона суммирования независимых средних квадратических отклонений:

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}; \tag{2.71}$$

при $r_{1,2}=\pm 1$

$$\sigma = \sqrt{\sigma_1^2 + 2r_{1,2}\sigma_1\sigma_2 + \sigma_2^2} = \sigma_1 \pm \sigma_2, \tag{2.72}$$

т. е. закон суммирования становится алгебраическим. Так, если

$\sigma_1 = 0,5$; $\sigma_2 = 0,2$, то $\sigma = 0,7$ при $r_{1,2}=1$ и $\sigma = 0,3$ при $r_{1,2} = -1$.

Для n случайных погрешностей законы суммирования принимают вид:

$$\sigma = \sqrt{\sum_{i=1}^n \sigma_i^2} \quad \text{для независимых величин} \quad (2.73)$$

и

$$\sigma = \sqrt{\sum_{i=1}^n \sigma_i^2 + 2 \sum_{i \neq j} r_{i,j} \sigma_i \sigma_j} \quad \text{для зависимых величин} \quad (2.74)$$

Порядок выполнения операций, предшествующих определению суммарной погрешности, был изложен в 2. 1.7.

2.1.9. Определение законов распределения случайных величин

Законы распределения случайных величин могут быть определены только экспериментально, на основании статистических данных. Теоретические законы распределения, которые широко используются в теории вероятностей, получены путем обработки большого числа наблюдений над случайными величинами и аналитического описания результатов, получаемых при обработке. Установление аналитической формы законов распределения случайных величин, а следовательно, и их числовых характеристик позволяет теоретически решать целый ряд задач. Некоторые из этих задач были рассмотрены выше применительно к погрешностям измерений. При этом предполагалось, что законы распределения отдельных погрешностей известны. Однако известны они только в том случае, если предварительно были поставлены эксперименты по наблюдению за величинами этих погрешностей, позволившие получить статистический материал, достаточный для установления законов их распределений. Таким образом, если принимается, например, что закон распределения погрешности отсчета в данном конкретном измерении есть закон равномерной плотности, то делается это на основании многочисленных экспериментов по определению законов распределения подобного вида погрешностей и, строго говоря, это решение требует проверки.

Не все виды составляющих погрешностей измерения достаточно хорошо изучены. Объясняется это тем, что поставить эксперимент так, чтобы результаты наблюдений, их разброс определялись только изучаемой погрешностью, очень трудно, а часто и невозможно. Поэтому часто законы распределений отдельных погрешностей устанавливаются на основании логических рассуждений. Законы распределения полной погрешности измерения, погрешности измерительного устройства всегда могут быть установлены и проверены при наличии необходимой статистики.

Если вид функции распределения погрешности известен, то эксперимент ставится для определения числовых характеристик. Эту задачу можно решить и при небольшом числе опытов. Конечно, степень приближения получаемых при этом оценок числовых характеристик к самим числовым характеристикам будет разной в зависимости от числа опытов.

Допустим, определяется закон распределении погрешности измерения величины X . При этом предполагается, что измеряемая величина неизменна за время эксперимента, а наблюдаемый разброс определяется погрешностью измерения, в том числе и погрешностью измерительной аппаратуры. Действия производятся в следующем порядке.

1. Производится не менее 100—200 наблюдений над измеряемой величиной в одинаковых условиях. Совокупность наблюдаемых значений дает простой статистический ряд, или простую статистическую совокупность. Пусть это будут значения x_1, x_2, \dots, x_n

2. Определяется среднее арифметическое из наблюдаемых значений x_i :

$$m_X^* = \frac{\sum_{i=1}^n x_i}{n}, \quad (2.75)$$

где звездочка означает, что значение m_X получено из опытных данных.

3. Определяются погрешности отдельных измерений

$$\delta_i = x_i - m_X^*. \quad (2.76)$$

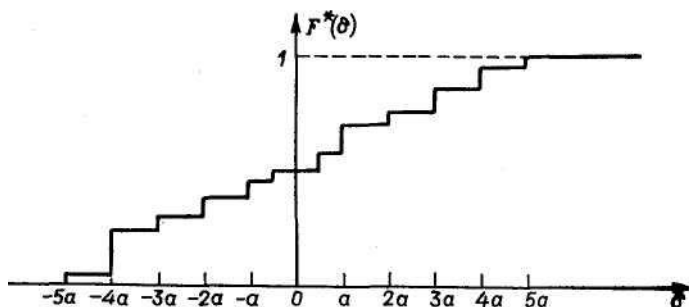


Рис. 2.15. Построение статистической функции распределения.

Случайные погрешности δ_i также образуют простой статистический ряд (первичную форму записи). Простой статистический ряд может быть непосредственно использован для определения статистической

функции распределения (плотности распределения), как это сделано в пп. 4 и 5, или может быть предварительно подвергнут обработке с целью получения статистического ряда (см. пп. 1а, 2а, 3а и ниже).

4. При построении статистической функции распределения находится наименьшее значение случайной погрешности. Пусть это будет $\delta_8 = -5a$ (рис. 2.15) и пусть такой величины погрешность только одна, $s_1 = 1$. Тогда частота ее появления при $n = 100$ равна $s_1/n = 1/100$, а $F^*(-5a) = 0$. Следовательно, в точке $-5a$ $F^*(\delta)$ имеет скачок, равный $1/100$.

5. Подсчитывается число остаточных погрешностей, меньших $-4a$, но больших $-5a$. Пусть их будет 5 ($\delta_6, \delta_{12}, \delta_{99}, \delta_{15}, \delta_{20}$), $s_2 = 5$. Тогда $F^*(-4a) = 1/100$, и в точке $-4a$ $F^*(\delta)$ делает скачек $5/100$. Делая аналогичные подсчеты числа погрешностей, лежащих в других интервалах, найдем статистическую функцию распределения $F^*(\delta)$. При $n \rightarrow \infty$ и уменьшении интервалов частота появления погрешностей приближается к вероятности их появления, а ступенчатая кривая $F^*(\delta)$ — к плавной кривой — функции распределения $F(\delta)$.

Построение статистического ряда проводится в следующем порядке.

1а. Выбирается интервал. Для этого в простом статистическом ряду находят наименьшую и наибольшую остаточные погрешности. Например, $\delta_{11} = 0,029$ и $\delta_{41} = 0,03$. Тогда диапазон наблюдаемых погрешностей $c = 0,06$. Число интервалов r принимают обычно равным 10—20. Пусть $r = 10$. Тогда интервал $a = s/r = 0,006$. Если c не делится на r точно, то частное округляют до 1-2 значащих цифр.

2а. Строится таблица, называемая статистическим рядом:

Интервалы	0— a		$a-2a$		$2a-3a$...	$(j-1)a - ja$...	$(r-1)a - ra$	
Частоты	+	-	+	-	+	-	+	-	+	-	+	-
$s_j/n = P_j^*$												

3а. Подсчитывается число погрешностей, лежащих в каждом интервале, отдельно с плюсом и минусом. В нижнюю строку таблицы записываются частоты их появления, определяемые как отношение числа погрешностей S_j , лежащих в данном интервале, к общему числу наблюдаемых погрешностей n . Погрешности, точно совпадающие по значению с границей интервала, могут быть отнесены либо к $(j-1)$ -интервалу, либо к j -интервалу. Если таких погрешностей, например, две, то целесообразно их разделить между смежными интервалами.

Статистический ряд можно оформить в виде гистограммы, как это сделано на рис. 2.16.

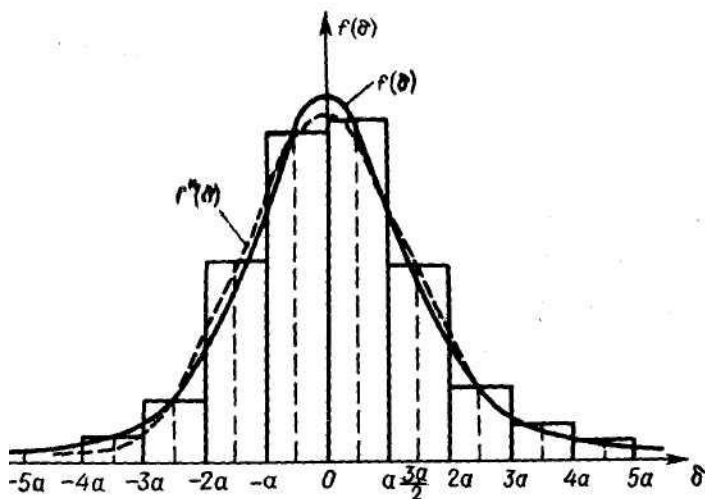


Рис. 2.16. Построение гистограммы.

Для этого по оси абсцисс откладываются интервалы $a, 2a, \dots, ra$. На каждом интервале как на основании строится прямоугольник, площадь которого равна частоте появления погрешностей, лежащих в данном интервале. Для определения высоты прямоугольника нужно его частоту разделить на величину интервала a . Тогда высоты (ординаты) равны P^*/a , т. е. пропорциональны частотам. Площадь гистограммы равна единице (из построения). При увеличении числа экспериментов, т. е. при увеличении n и уменьшении величины интервалов, гистограмма все больше будет приближаться к кривой, ограничивающей площадь, равную единице. Таким образом, гистограмма дает статистическую плотность распределения $f^*(\delta)$ погрешностей, приближающуюся к теоретической $f(\delta)$ при $n \rightarrow \infty$. Поэтому ось ординат гистограммы можно обозначить через $f^*(\delta)$. Очевидно, что по статистическому ряду, представленному в любой форме, можно легко построить и статистическую функцию распределения. На основании полученного статистического закона распределения решается вопрос, какому теоретическому закону распределения он соответствует. Установление этого положения возможно лишь с некоторым приближением вследствие ограниченного числа экспериментов, и поэтому гипотеза о согласованности статистического и предполагаемого теоретического распределений требует проверки.

Использование статистического ряда для определения закона распределения и числовых характеристик, о которых речь пойдет ниже, правомерно, если статистические данные были получены в один и тот же момент времени или хотя бы в течение очень короткого промежутка времени для одного и того же значения X . Если же эксперимент растягивается на длительное время или проводится с интервалами (часть, например, сегодня, часть — завтра), т. е. группами, или совокупность получена путем объединения групп измерений разных значений X , то прежде чем производить объединение групп в один статистический ряд необходимо проверить допустимость (обоснованность) различий в средних арифметических m_X групп измерений, а также их равномерность.

2.1.10. Числовые характеристики статистического распределения

Статистическое распределение имеет числовые характеристики, аналогичные числовым характеристикам теоретического распределения.

Аналогом математического ожидания случайной величины является среднее арифметическое. Если X — случайная величина, то

$$M^* [X] = m_X^* = \frac{\sum_{i=1}^n x_i}{n}, \quad (2.77)$$

где x_i — значения, принимаемые X в опытах; n — число опытов.

Среднее арифметическое, обычно называемое статистическим средним, при большом n является наиболее достоверным значением измеряемой величины; его следует принять за действительное значение X_0 как значение, наиболее приближающееся к истинному X_{00} .

Одной из важнейших характеристик случайных величин является дисперсия — второй центральный момент, равный математическому ожиданию квадрата центрированной случайной величины:

$$D [\hat{X}] = M [(X - m_X)^2] = D [\delta] = M [\delta^2] \quad (2.78)$$

(дисперсия X и есть дисперсия δ).

Для определения статистической дисперсии математическое ожидание в (2.78) заменим его статистической аналогией — средним арифметическим. Тогда статистическая дисперсия

$$D^* [\hat{X}] = D^* [\delta] = \frac{\sum_{i=1}^n (x_i - m_X^*)^2}{n}, \quad (2.79)$$

статистическое среднее квадратическое отклонение

$$\sigma^* = \sigma_X^* = \sqrt{\frac{\sum_{i=1}^n (\delta_i)^2}{n}}. \quad (2.80)$$

Аналогично можно найти выражения для статистических начальных и центральных моментов:

$$\alpha_s^*[X] = \frac{\sum_{i=1}^n x_i^s}{n}, \quad (2.81)$$

$$\mu_s^*[X] = \mu_s^*[\delta] = \frac{\sum_{i=1}^n \delta_i^s}{n}. \quad (2.82)$$

В заключение следует отметить, что

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^n \delta_i^2}{n}} \quad (2.83)$$

является средним квадратическим отклонением случайных погрешностей δ_i от математического ожидания, равного нулю, и, следовательно, является числовой характеристикой точности каждого отдельного измерения в данной серии опытов.

Числовой характеристикой погрешности результата, т. е. погрешности установления действительного значения X_0 измеряемой величины как степени отличия его от истинного значения, является среднее квадратическое отклонение результата измерения

$$\sigma_m^* = \sigma^* / \sqrt{n}. \quad (2.84)$$

Действительно,

$$D[m_X^*] = D\left[\sum_{i=1}^n \frac{x_i}{n}\right]. \quad (2.85)$$

В дальнейшем индекс x у σ и D опущен, если речь идет δ . Индекс m относится к математическому ожиданию m_X .

Так как опыты независимы, то дисперсия суммы равна сумме дисперсий, т.е.

$$\begin{aligned} D\left[\sum_{i=1}^n \frac{x_i}{n}\right] &= \sum_{i=1}^n D\left[\frac{x_i}{n}\right] = \sum_{i=1}^n \frac{1}{n^2} D[x_i] = \\ &= \sum_{i=1}^n \frac{1}{n^2} M[(x_i - m_X^*)^2]. \end{aligned} \quad (2.86)$$

Заменяя операцию математического ожидания определением среднего арифметического, получим

$$\begin{aligned}
 D [m_X^*] &= \sum_{i=1}^n \frac{1}{n^2} \frac{\sum_{i=1}^n (x_i - m_X^*)^2}{n} = \\
 &= D^* [X[\sum_{i=1}^n \frac{1}{n^2} = \frac{1}{n} D^* [X] = \frac{1}{n} D_X^*.
 \end{aligned}
 \tag{2.87}$$

Таким образом, определяя X по среднему статистическому, мы допускаем ошибку, в \sqrt{n} раз меньшую, чем при определении по результату одного измерения.

Производство большого числа опытов в одинаковых условиях, естественно, затруднительно и требует большого времени. Статистические же данные, получаемые при этом, часто оказываются малоэффективными и даже неправильными вследствие изменения измеряемой величины и условий измерения. Поэтому названное выше число опытов (100—200) делается только при определении законов распределения погрешностей. При определении же числовых характеристик, что, понятно, приходится делать много чаще, достаточно 10—20 опытов.

Числовые характеристики определенные на основе ограниченной статистической совокупности, всегда содержат элемент случайности, т. е. сами являются случайными величинами, оценками числовых характеристик. Если оценивается параметр a , то его оценка \hat{a} . Оценки, определяемые по результатам ограниченного числа наблюдений, должны вычисляться по таким формулам, чтобы их можно было принять за состоятельные, чтобы использование оценок по крайней мере не приводило к изменению (смещению) величины X или систематической погрешности θ (как математического ожидания случайной погрешности Δ); чтобы найденная несмещенная оценка, как случайная величина, обладала наименьшей дисперсией по сравнению с другими возможными оценками (была бы наиболее эффективной).

Поясним введенные понятия подробнее.

Оценка \hat{a} параметра (числовой характеристики) a называется состоятельной, если при увеличении числа опытов n она приближается, сходится по вероятности к параметру a .

Оценка \hat{a} называется несмещенной, если при использовании ее вместо a не происходит увеличения или уменьшения математического ожидания, т. е.

$$M [\hat{a}] = a.
 \tag{2.88}$$

Несмещенная оценка называется эффективной, если она обладает по сравнению с другими несмещенными оценками наименьшей дисперсией, т. е.

$$D[\hat{a}] = \min, \quad (2.89)$$

Удовлетворить всем этим требованиям при выборе оценки принципиально можно. Но формулы, которыми при этом приходится пользоваться, оказываются слишком сложными, чтобы их можно было использовать в широкой практике. Поэтому обычно используют для расчета оценок более простые зависимости, хотя они и приводят к получению оценок, несколько отличных от эффективных или несмещенных. Рассмотрим оценки некоторых числовых характеристик.

В качестве оценки математического ожидания измеряемой величины естественно принять среднее арифметическое

$$\hat{m}_X = \frac{\sum_{i=1}^n x_i}{n} = m_X^*, \quad (2.90)$$

которое при увеличении n сходится по вероятности к m_X (формулы для определения \hat{m}_X и m_X^* одинаковы, но значения их будут несколько разными). Так как эта оценка является состоятельной, то она будет и несмещенной, ибо

$$M[\hat{m}_X] = M[m_X] = m_X. \quad (2.91)$$

Дисперсия этой оценки [см. аналогичную формулу (2.87) для $D(m_X^*)$].

$$D[\hat{m}_X] = \frac{1}{n} D_X. \quad (2.92)$$

Доказать, является ли эта оценка эффективной, можно, если X распределена по нормальному закону. Для определения $D[\hat{m}_X]$ необходимо найти оценку $D_X - \hat{D}_X$. Выберем в качестве \hat{D}_X статистическую дисперсию D_X^* . Тогда

$$\hat{D}_X = D_X^* = \frac{\sum_{i=1}^n (x_i - \hat{m}_X)^2}{n}. \quad (2.93)$$

Проверим прежде, является ли она состоятельной. Учитывая независимость опытов, можно D_X^* представить в виде

$$D_X^* = \frac{\sum_{i=1}^n x_i^2}{n} - \hat{m}_X^2. \quad (2.94)$$

Первый член есть среднее арифметическое из n значений X^2 . По вероятности он сходится к второму начальному моменту $\alpha_2[X]=M[X^2]$. Второй член по вероятности сходится к m^2_X . Тогда D^*_X сходится по вероятности к величине

$$D_X = \alpha_2 [X] - m^2_X, \quad (2.95)$$

т. е. принятая оценка D_X (2.93) является состоятельной.

Для определения несмещенности D^*_X как оценки D_X в (2.94)

вместо \hat{m}_X подставим его значение:

$$\begin{aligned} \hat{D}_X = D_X^* &= \frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2 = \frac{n-1}{n^2} \sum_{i=1}^n x_i^2 - \\ &- 2 \frac{\sum_{i \neq j} x_i x_j}{n^2} \end{aligned} \quad (2.96)$$

и найдем математическое ожидание оценки

$$M[\hat{D}_X] = \frac{n-1}{n^2} \sum_{i=1}^n M[X_i^2] - \frac{2}{n^2} \sum_{i \neq j} M[X_i X_j]. \quad (2.97)$$

Так как дисперсия D^*_X не зависит от того, чему равно математическое ожидание, то примем его равным нулю, т. е. будем считать X_i центрированной величиной. Тогда

$$M[X_i^2] = M[\hat{X}_i^2] = D_X, \quad (2.98)$$

$$\sum_{i=1}^n M[X_i^2] = nD_X, \quad (2.99)$$

$$M[X_i X_j] = M[\hat{X}_i \hat{X}_j] = K_{i,j} = 0, \quad (2.100)$$

$$M[\hat{D}_X] = \frac{n-1}{n} D_X. \quad (2.101)$$

Таким образом, принятая оценка D_X или D^*_X , не является несмещенной, так как ее математическое ожидание несколько меньше D_X . Определяя по (2.93) дисперсию X или принимая в качестве оценки погрешности δ_i отдельного измерения $\sqrt{D^*_X} = \sqrt{D^*}$, мы допускаем систематическую ошибку, приводящую к получению несколько заниженных численных значений дисперсии. Если в качестве оценки D_X принять

$$\hat{D}_X = \frac{n}{n-1} D_X^* = \frac{\sum_{i=1}^n (x_i - \hat{m}_X)^2}{n-1}, \quad (2.102)$$

$$\text{то } M[\hat{D}_X] = M\left[\frac{\sum_{i=1}^n (x_i - \hat{m}_X)^2}{n(n-1)}\right] = D_X. \quad (2.103)$$

Полученная оценка (2.102) является также состоятельной. Подставив ее в (2.92), найдем дисперсию оценки математического ожидания (действительного значения измеряемой величины)

$$D_m = \frac{1}{n} \hat{D}_X = \frac{\sum_{i=1}^n (x_i - \hat{m}_X)^2}{n(n-1)}. \quad (2.104)$$

Соответственно для средних квадратических отклонений отдельных погрешностей и погрешности результата получим оценки:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (x_i - \hat{m}_X)^2}{(n-1)}} \quad (\text{формула Бесселя}), \quad (2.105)$$

$$\hat{\sigma}_m = \hat{\sigma}_{X_0} = \sqrt{\frac{\sum_{i=1}^n (x_i - \hat{m}_X)^2}{n(n-1)}}. \quad (2.106)$$

При определении числовых характеристик всегда пользуются ограниченным рядом наблюдений. Поэтому после обработки данных наблюдений получают лишь их оценки; чем больше число наблюдений, тем ближе оценки (если они состоятельные) к самим числовым характеристикам.

Выше было показано, что вероятность появления погрешностей, лежащих в разных интервалах, не одинакова. Так, по приложению 1 можно найти, что погрешность измерения не превышает $\pm 2\sigma_m$ с вероятностью 0,95; с вероятностью 0,997 погрешность измерения не превысит $\pm 3\sigma_m$ и т. д. Таблица составлена на основании теоретических расчетов и поэтому может быть использована для определения теоретических и статистических (при $n \rightarrow \infty$) числовых характеристик. Определять оценки по этой таблице было бы неверно.

Для определения оценок числовых характеристик существуют таблицы Стьюдента — Фишера, определяющие связь между погрешностями измерения и вероятностью их появления для различного числа наблюдений, т. е. закон распределения случайных ошибок с учетом количества статистического материала. Используя таблицы Стьюдента — Фишера, можно проследить деформацию нормального закона. При большом объеме статистических

наблюдений n статистический закон распределения и его числовые характеристики, как было показано выше, совпадают с теоретическими. Действительно, данные последней строки приложения 2 совпадают с данными приложения 1. При уменьшении числа измерений вероятности появления погрешностей тех же значений уменьшаются. Так, при $n=10$ вероятное появления погрешностей, меньших $\pm 2\sigma_m$, становится равной 0,9266, а не 0,9545, как это имеет место для теоретического и статистического законов распределения (при $n \rightarrow \infty$).

Вероятность появления погрешностей, меньших $\pm z\sigma$, согласно распределению Стьюдента — Фишера,

$$P = 2 \int_0^{z^2} f_C(n, z) dz, \tag{2.107}$$

где $f_C(n, z)$ - плотность распределения Стьюдента — Фишера
Плотность распределении $f_C(n, z)$ выражается зависимостью

$$f_C(n, z) = \frac{\left(\frac{n}{2}\right)!}{\left(\frac{n-1}{2}\right)! \sqrt{\pi(n-1)}} \left[1 + \frac{z^2}{n-1}\right]^{-\frac{n}{2}}. \tag{2.108}$$

Таблицы Стьюдента- Фишера позволяют решить и обратную задачу — выбор числа экспериментов для обеспечения заданной степени приближения оценок к статистическим характеристикам. При $n=121$ погрешность в определении вероятности появления погрешностей, лежащих в разных интервалах, составляет 0,5-2%.

В заключение следует отметить, что числовые характеристики являются точечными параметрами законов распределения. Неверной является фраза «вероятность появления погрешности 2σ равна 0,4773», так как вероятность появления погрешности, точно равной 2σ , равна нулю. $P = 0,4773$ есть вероятность появления погрешностей, лежащих в интервале от 0 до 2σ .

Аналогично результат измерения X (истинное значение) не может быть выражен одним числом, так как при обработке статистических данных были получены лишь математическое ожидание $m_X = X_0$, которое может быть принято за действительное значение, и среднее квадратическое отклонение σ_m . Следовательно, истинное значение лежит в некотором интервале. Ширина интервала зависит от требуемой вероятности его определения, или, как принято говорить, от заданной доверительной вероятности. Так, при нормальном законе распределения X и заданной доверительной вероятности $P = 0,997$

$$X_{00} = \hat{m}_X \pm 3\hat{\sigma}_m, \quad (2.109)$$

т. е. истинное значение лежит в интервале

$$(\hat{m}_X + 3\hat{\sigma}_m) \div (\hat{m}_X - 3\hat{\sigma}_m).$$

Интервал, в котором лежит истинное значение измеряемой величины с заданной доверительной вероятностью, называется **доверительным интервалом**.

Доверительные интервалы для погрешностей измерения могут задаваться различно; им будут соответствовать определенные значения доверительных вероятностей. Так, доверительному интервалу $\pm z\sigma$ соответствует доверительная вероятность (при нормальном законе распределения) появления погрешностей в этом интервале

$$P[-z\sigma < \delta \leq z\sigma] = \frac{2}{\sqrt{2\pi}} \int_0^{z\sigma} \exp\left(-\frac{z^2}{2}\right) dz. \quad (2.110)$$

Доверительному интервалу $z\sigma - (z+1)\sigma$ соответствует вероятность появления погрешностей, лежащих в этом интервале:

$$\begin{aligned} P[z\sigma < \sigma \leq (z+1)\sigma] &= \frac{1}{\sqrt{2\pi}} \int_{z\sigma}^{(z+1)\sigma} \exp\left(-\frac{z^2}{2}\right) dz = \\ &= \frac{1}{\sqrt{2\pi}} \int_0^{(z+1)\sigma} \exp\left(-\frac{z^2}{2}\right) dz - \frac{1}{\sqrt{2\pi}} \int_0^{z\sigma} \exp\left(-\frac{z^2}{2}\right) dz. \end{aligned} \quad (2.111)$$

При построении интервалов по оценкам следует пользоваться распределением Стьюдента — Фишера.

Оценки числовых характеристик являются случайными величинами, так как они получены из ограниченной выборки из результатов измерений. Для оценок, как случайных величин, также можно найти доверительные интервалы.

На рис. 2.17 самым узким прямоугольником выделен интервал, в котором лежит истинное значение измеряемой величины, определенный по математическому ожиданию \hat{m}_X и $\hat{\sigma}_m$.



Рис. 2.17. Уточнение интервала определения X.

Так как $\hat{\sigma}_m$, как случайная величина, лежит в некотором интервале (на рис. 2.16 более высокие прямоугольники), то истинное значение измеряемой величины лежит для заданной доверительной вероятности в более широком интервале. При нормальном законе распределения X отношение

$$v = \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \quad (2.112)$$

имеет так называемое хи-квадрат распределение Пирсона с $k = n - 1$ степенями свободы. Вид кривой плотности χ^2 -распределения изображен на рис. 2.18.

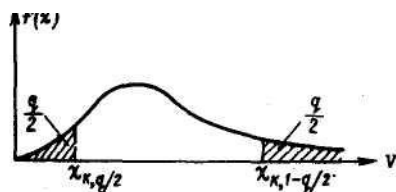


Рис. 2.18. Кривая плотности χ^2 -распределения.

Значения χ^2_{kP} , соответствующие вероятностям P того, что при $n=k+1$ отношение v будет меньше χ^2_{kP} , приведены в приложении 3. Доверительный интервал строится таким образом, чтобы вероятность выхода дисперсии за его границы не превышала некоторой малой величины, например q . Примем вероятности выхода за каждую границу одинаковыми, равными $0,5 q$ (на рис. 2.18 это заштрихованные площади). Вероятности, соответствующие этим границам,

$$P_n = \frac{1}{2} q \text{ и } P_v = 1 - \frac{1}{2} q, \quad (2.113)$$

а значения χ^2 , полученные из приложения 3, соответственно равны χ_n и χ_v .

Вероятность того, что отношение v не выйдет за пределы интервала, ограниченного χ_v^2 и χ_n^2 ,

$$\begin{aligned}
 P[\chi_{\text{H}}^2 < \sigma \leq \chi_{\text{B}}^2] &= P\left[\chi_{\text{H}}^2 < \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq \chi_{\text{B}}^2\right] = \\
 &= P\left[\frac{(n-1)\hat{\sigma}^2}{\chi_{\text{H}}^2} > \sigma^2 \geq \frac{(n-1)\hat{\sigma}^2}{\chi_{\text{B}}^2}\right] = 1 - q,
 \end{aligned}
 \tag{2.114}$$

или для среднего квадратического отклонения

$$P\left[\frac{\sqrt{n-1}\hat{\sigma}}{\chi_{\text{H}}} > \sigma \geq \frac{\sqrt{n-1}\hat{\sigma}}{\chi_{\text{B}}}\right] = 1 - q.
 \tag{2.115}$$

Доверительный интервал, таким образом, определяется границами

$$\hat{\sigma}_{\text{H}} = \frac{\sqrt{n-1}\hat{\sigma}}{\chi_{\text{H}}} \quad \text{и} \quad \hat{\sigma}_{\text{B}} = \frac{\sqrt{n-1}\hat{\sigma}}{\chi_{\text{B}}}.
 \tag{2.116}$$

Этот интервал на рис. 2.16 отмечен самым высоким прямоугольником.

Итак, интервал, в котором лежит истинное значение измеряемой величины, при ограниченном числе измерений уточняется дважды. Первое уточнение — определение оценок числовых характеристик с учетом числа опытов, т. е. из распределения Стьюдента—Фишера.

Фактически это позволяет уточнить $\hat{\sigma}_m$. Второе уточнение связано с установлением вероятности отклонения от принятой оценки, которое только что было рассмотрено. Следует отметить, что при достаточно больших n эти уточнения часто не производятся.

2.1.11. Проверка правдоподобия гипотезы о соответствии статистического закона теоретическому

В 2.1.9 была изложена методика определения гистограммы распределения, т. е. плотности распределения вероятностей. Естественно, что статистическое распределение несет элемент случайности, связанный прежде всего с ограничением объема выборки из обширной совокупности значений (так называемой генеральной совокупности), которые может принять измеряемая величина X и, следовательно, погрешность $\delta = X - X_{00}$. Поэтому статистическая кривая распределения может принимать форму, на первый взгляд не отвечающую ни одному известному, теоретически описанному, закону распределения. Однако знание характера изучаемого явления (погрешности) и внимательное рассмотрение статистической кривой

в большинстве случаев позволяют подобрать теоретический закон, графическое изображение плотности распределения которого близко к подобной кривой статистического распределения (за исключением некоторых участков, где проявляются элементы случайности). Так, на рис. 2.16 гистограмма несколько неодинакова по обе стороны от оси ординат; в этом проявляется здесь элемент случайности. Но даже с первого взгляда можно сказать, что ей соответствует нормальный закон распределения, но с неизвестными числовыми характеристиками. Поэтому следующей задачей, которую приходится решать, является задача выравнивания статистических рядов, заключающаяся в рациональном выборе числовых характеристик и параметров, при котором соответствие между теоретическими и статистическими распределениями будет наиболее полным. Для нормального распределения такими характеристиками будут математическое ожидание и среднее квадратическое отклонение, так как

$$f(x) = f(\delta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - m_x)^2}{2\sigma^2}\right]. \quad (2.117)$$

Для выравнивания статистических рядов можно использовать метод моментов, заключающийся в выборе теоретических σ и m , совпадающих со статистическими σ^* и m^* . Таким образом, если распределение характеризуется двумя параметрами, ставится условие совпадения первых двух моментов; если $f(x)$ зависит от трех параметров, то должны совпадать первые три момента, и т. д. Возможны и другие методы выравнивания статистических рядов.

Как бы тщательно ни была подобрана теоретическая кривая, расхождения между нею и статистической кривой все-таки будут, причем вызваны они могут быть как случайными обстоятельствами, так и тем, что выбранный теоретический закон плохо выравнивает статистическое распределение и потому должен быть заменен другим. Для проверки правдоподобия гипотез используются те или иные критерии согласия. Чаще всего применяется критерий согласия Пирсона, критерий χ^2 -

Порядок выполнения операций после построения гистограммы (см. 2.1.9) может быть следующим.

1. Для средних значений погрешностей в каждом интервале определяется их теоретическая и статистическая вероятность. В интервале от 0 до a среднее значение погрешности равно $a/2$. В интервале от a до $2a$ оно равно $a + a/2 = 3/2a$, в интервале от $(j-1)a$ до ja

оно будет $(j-1)a + \frac{a}{2} = \frac{2j-a}{2}$ и т. д. (см. таблицу в 2.1.9 и рис. 2.16). Для этих значений погрешностей определяются соответствующие им вероятности путем умножения ординаты теоретического закона в этой точке на длину интервала:

$$P_j = f_j(\delta) a. \quad (2.118)$$

Значения P_j^* берут из таблицы в 2.1.9.

2. В качестве меры расхождения между теоретическим и статистическим распределением принимается сумма квадратов отклонений $(P_j^* - P_j)$, взятых с некоторыми весами C_j :

$$v = \sum_{j=1}^r C_j (P_j^* - P_j)^2. \quad (2.119)$$

Необходимость введения весового коэффициента связана с тем, что $P_j^* - P_j$ для разных j неравноценна по значимости. Если разность мала, а P_j велико, то разностью в P_j^* и P_j можно пренебречь; если же P_j мало, то даже маленькая разность между P_j^* и P_j становится значительной. Если в качестве C_j , как предложил Пирсон, выбрать n/P_j , то при больших n закон распределения величины

$$v = n \sum_{j=1}^r \frac{(P_j^* - P_j)^2}{P_j} \quad (2.120)$$

приближается к закону распределения χ^2 .

3. Определяется число степеней свободы χ^2 -распределения. При требовании совпадения теоретического и статистического распределений число степеней свободы $k = n - 3$, так как накладываются три условия:

$$\sum_{j=1}^r P_j = 1; \quad \sum_{j=1}^r \bar{\delta}_j P_j^* = m_{\delta}^* = 0; \quad \sum_{j=1}^r \delta_j^2 P_j^* = D_{\delta}^*, \quad (2.121)$$

где $\bar{\delta}_j$ — средние значения погрешностей в интервалах.

4. Задаваясь вероятностью q выхода величины v за границы интервала, находят вероятности, соответствующие верхней и нижней границам:

$$P_n = \frac{1}{2} q \quad \text{и} \quad P_v = 1 - \frac{1}{2} q. \quad (2.122)$$

По приложению 3 для степени свободы $k = n - 3$ находят χ_{n}^2 и χ_{v}^2 . Если

$$\chi_n^2 < v \leq \chi_v^2, \quad (2.123)$$

то расхождения в законах определяются чисто случайными причинами и ограниченностью числа измерений n , и гипотезу о совпадении теоретического и статистического законов можно считать правдоподобной.

Можно поступить несколько иначе: подсчитав $\nu = \chi^2$ и $k = n-3$, по приложению 3 определить вероятность того, что за счет чисто случайных причин мера расхождения ν будет не меньше, чем активно наблюдаемое в данной серии экспериментов значение χ^2 . Если эта вероятность весьма мала, то принятую гипотезу о том, что выбранный теоретический закон есть закон распределения X , δ , следует отвергнуть. Наоборот, если эта вероятность велика, то гипотезу следует принять как правдоподобную.

2.1.12. Погрешности косвенного измерения

Косвенные измерения — это такие, при которых значение величины определяется не непосредственно, а по значениям других величин, измеренных прямым методом и связанных с определяемой величиной известной зависимостью. Пусть искомая величина определяется выражением

$$Y = \varphi(X_1, X_2, \dots, X_i, \dots, X_m), \quad (2.124)$$

где $X_1, X_2, \dots, X_i, \dots, X_m$ — величины, определяющие Y и подвергаемые прямым измерениям.

Для упрощения положим, что погрешности измерения $X_1, X_2, X_3, \dots, X_m$ не коррелированы. Так обычно и бывает на практике, если погрешность не вызвана каким-либо общим фактором, изменяющимся случайным образом. Таким фактором чаще всего является температура. Однако при постановке экспериментов по накоплению статистического материала для определения числовых характеристик за изменением температуры, если ее не удастся поддерживать постоянной, следят и в результаты измерений вносят поправки. При косвенных измерениях, кроме того, всегда есть возможность сделать $X_1, X_2, X_3, \dots, X_i, \dots, X_m$ независимыми, производя n их измерений в разное время (например, сначала n раз измеряется X_1 , а затем X_2 и т. д.). Если систематические погрешности измерения X_1, X_2, \dots, X_m исключены и число опытов при измерении каждой из этих величин равнялось n , то

$$X_i = m_{X_i} \pm z_i \sigma_{mi} = m_{X_i} \pm \lambda_i, \quad (2.125)$$

где m_{X_i} — математическое ожидание X_i , равное его среднему арифметическому;

$\sigma_{mi} = \frac{\sigma_i}{\sqrt{n}}$ — среднее квадратическое отклонение m_{Xi} ; σ_i — среднее квадратическое отклонение X_i или среднее квадратическое отклонение одного (отдельного) измерения; z_i — коэффициент, определяющий доверительный интервал, при заданной вероятности определения X_i ; $\lambda_i = z_i \sigma_{mi}$.

Погрешность измерения Y не равна сумме погрешностей измерения X_1, X_2, \dots, X_m — $\lambda_1, \lambda_2, \dots, \lambda_m$, так как Y связано с определяющими его величинами функциональной зависимостью.

Результат измерения Y должен быть выражен через числовые характеристики, т. е. представлен в виде

$$Y = m_Y \pm z \sigma_{mY} \approx m_Y \pm \lambda, \tag{2.126}$$

где m_Y — математическое ожидание Y ; σ_{mY} — среднее квадратическое отклонение m_Y ; λ — погрешность измерения Y , зависящая от принятой доверительной вероятности,

Разложим правую часть (2.124) в ряд Тейлора, сохранив в разложении только первые два члена. Тогда

$$Y = \varphi(X_1, X_2, \dots, X_i, \dots, X_m) \approx \varphi(m_{X1}, m_{X2}, \dots, m_{Xi}, \dots, m_{Xm}) + \frac{\partial \varphi}{\partial X_1} (X_1 - m_{X1}) + \frac{\partial \varphi}{\partial X_2} (X_2 - m_{X2}) + \dots + \frac{\partial \varphi}{\partial X_m} (X_m - m_{Xm}). \tag{2.127}$$

Учитывая, что $X_i - m_{Xi} = \delta_i$, перепишем (2.127) в виде

$$Y = \varphi(m_{X1}, m_{X2}, \dots, m_{Xm}) + \sum_{i=1}^m \frac{\partial \varphi}{\partial X_i} \delta_i. \tag{2.128}$$

Функция (2.128) является линейной. Числовые характеристики ее:

математическое ожидание

$$m_Y = \varphi(m_{X1}, m_{X2}, \dots, m_{Xm}), \tag{2.129}$$

дисперсия

$$D_{mY} = M[Y^2] = \left(\frac{\partial \varphi}{\partial X_1} \sigma_{m1}\right)^2 + \left(\frac{\partial \varphi}{\partial X_2} \sigma_{m2}\right)^2 + \dots + \left(\frac{\partial \varphi}{\partial X_m} \sigma_{mm}\right)^2 \tag{2.130}$$

и среднее квадратическое отклонение Y от m_Y

$$\sigma_{mY} = \sqrt{\left(\frac{\partial \varphi}{\partial X_1} \sigma_{m1}\right)^2 + \left(\frac{\partial \varphi}{\partial X_2} \sigma_{m2}\right)^2 + \dots + \left(\frac{\partial \varphi}{\partial X_m} \sigma_{mm}\right)^2}. \tag{2.131}$$

Для заданной доверительной вероятности погрешность

измерения Y не превысит $z\sigma_{mY} = \lambda$.

Выражения вида $\frac{\partial \varphi}{\partial X_i} \sigma_{mi}$ называются коэффициентами влияния, а

$$\left(\frac{\partial \varphi}{\partial X_i} \sigma_{mi} \right) \tag{2.132}$$

есть средние квадратические отклонения частных составляющих погрешностей косвенного измерения. Коэффициенты влияния определяются при значениях X , равных их математическим ожиданиям.

Следует отметить, что при однократном измерении X_1, X_2, \dots, X_m среднее квадратическое отклонение σ_Y можно рассчитать по формуле (2.131), заменив в ней $\sigma_{m1}, \sigma_{m2}, \dots, \sigma_{mm}$ на средние квадратические отклонения $\sigma_1, \sigma_2, \dots, \sigma_m$.

Результат однократного измерения Y определяется при этом по формуле (2.129) подстановкой в нее вместо математических ожиданий X_1, X_2, \dots, X_m результатов их однократных измерений.

Если $\lambda_1, \lambda_2, \dots, \lambda_i, \dots, \lambda_m$ коррелированы, например, воздействиями температуры и корреляционная матрица

$$\| K_{ij} \| = \begin{vmatrix} K_{11} & K_{12} & \dots & K_{1m} \\ & K_{22} & \dots & K_{2m} \\ & & \dots & \dots \\ & & & K_{mm} \end{vmatrix} \tag{2.133}$$

известна, то дисперсия Y

$$D_{mY} = \sum_{i=1}^m \left(\frac{\partial \varphi}{\partial X_i} \right)^2 D_{mi} + 2 \sum_{i \neq j} \left(\frac{\partial \varphi}{\partial X_i} \right) \left(\frac{\partial \varphi}{\partial X_j} \right) K_{ij}. \tag{2.134}$$

Так как

$$K_{ij} = r_{ij} \sigma_{mi} \sigma_{mj}, \tag{2.135}$$

то среднее квадратическое отклонение Y от его математического ожидания m_Y

$$\sigma_{mY} = \sqrt{\sum_{i=1}^m \left(\frac{\partial \varphi}{\partial X_i} \sigma_{mi} \right)^2 + 2r_{ij} \left(\frac{\partial \varphi}{\partial X_i} \right) \left(\frac{\partial \varphi}{\partial X_j} \right) \sum_{i \neq j} \sigma_{mi} \sigma_{mj}}. \tag{2.135}$$

Установление корреляционных связей между погрешностями в большинстве случаев затруднительно. Обычно, если они есть, то r_{ij} принимаются равными 1; если их нет, то $r_{ij} = 0$.

2.1.13. Промахи

Промахами называются большие случайные погрешности, явно искажающие результат измерения. Они не могут быть объяснены условиями эксперимента и возникают, как правило, вследствие оплошности наблюдателя: неправильного отсчета, неверной записи результата, неправильных манипуляций при проведении эксперимента. Иногда грубая погрешность связана с резким кратковременным изменением какого-либо внешнего фактора.

Наблюдения содержащие промахи, следует исключать из рассмотрения, если есть уверенность в том, что погрешность этого наблюдения не есть крупная случайная погрешность, подчиняющаяся закону распределения наблюдаемой величины, хотя и редко встречающаяся. Чтобы решить, какая погрешность может быть принята за промах, вводится понятие **критерия грубых погрешностей**. Наиболее простым является критерий Райта, согласно которому погрешность $|\delta_i|_{\max}$, большую 4σ для нормального закона распределения при большом n , можно считать промахом, так как вероятность появления погрешностей (не промахов), больших 4σ , составляет очень маленькую величину — 0,000064.

При небольшом числе измерений лучше использовать критерий, величина которого может быть выбрана в зависимости от n и заданной вероятности, т. е. не 4σ , а $z_{np}\sigma$. Для определения

$$z_{np} = \frac{|\delta_i|_{\max}}{\sigma} \quad (2.137)$$

при различных P и n составлены таблицы (см. приложение 4). Определение промаха в этом случае производится путем сравнения

$$z = \frac{|\delta_i|_{\max}}{\sigma}, \quad (2.138)$$

полученного из опытных данных, с z_{np} .

Порядок обработки экспериментальных данных, если они предположительно содержат промах, несколько усложняется. Вначале, как обычно, по случайным погрешностям находят оценку среднего квадратического отклонения σ . Затем, выявив предположительно наблюдение, содержащее промах, для случайной погрешности, соответствующей этому наблюдению, находят z по (2.138) и сравнивают его либо с 4 (по Райту), либо с z_{np} , полученным из приложения 4 по заданной вероятности и числу измерений n . Если

$$z > z_{np}, \quad (2.139)$$

то наблюдение принимается за промах и исключается из статистического ряда. Полученный новый ряд обрабатывается в обычном порядке.

2.1.14. Систематические погрешности. Оценка неисключенного остатка.

Выше уже отмечалось, что практически всегда погрешность измерения является случайной величиной, характеризуемой математическим ожиданием и средним квадратическим отклонением:

$$\Delta = m_{\Delta} \pm z\sigma_m = \theta \pm z\sigma_m, \quad (2.140)$$

где $m_{\Delta} = \theta$ — математическое ожидание случайной величины Δ , называемое систематической погрешностью; σ_m — среднее квадратическое отклонение Δ от математического ожидания m_{Δ} ; $z\sigma_m$ — случайная составляющая погрешности измерения.

Действительно, чем бы ни вызывалась погрешность измерения, из чего бы она ни складывалась (из погрешностей от изменения влияющих факторов, методических, инструментальных, личных), всегда можно наблюдать, что она имеет некоторую систематическую и случайную составляющие. Так, трение в приборе вызывает погрешность в его показаниях и, следовательно, приводит к погрешности измерения, случайным образом изменяющейся около некоторого значения, принимаемого за математическое ожидание. Систематическую составляющую погрешности от трения исключают путем корректировки нуля, т. е. в одной точке шкалы. В других точках шкалы систематическая составляющая может иметь другое значение, поэтому корректировкой нуля она полностью не исключается.

Погрешность отсчета при наличии параллакса также будет иметь и систематическую, и случайную составляющие.

Постановку эксперимента, измерения стараются осуществить так, чтобы влияние факторов, вызывающих систематическую погрешность, было исключено. Для этого стабилизируется напряжение питания схемы, исключается влияние магнитных полей путем правильного расположения аппаратуры и т. д. Полностью исключить систематические погрешности, конечно, не удастся. На неисключенные систематические погрешности вносится поправка. Однако обнаружить присутствие систематической погрешности и определить ее значение легко, если она вызвана каким-либо одним, в крайнем случае двумя факторами. Например, температурой и магнитным полем. Так, наблюдая за изменениями температуры по показанию прибора при $X = \text{const}$, можно легко обнаружить и определить систематическую погрешность, вызванную изменением

температуры. Поправка, которую следует ввести в результат измерения, при изменении влияющего фактора с T_1 на T_2 может быть точно рассчитана. При повторении экспериментов условия проведения не остаются строго неизменными, и фактор может принимать значения $T_2 \pm \Delta T$, причем эти изменения ΔT обычно трудно обнаружить. Принимая, таким образом, в качестве поправки определенную с помощью специального эксперимента систематическую погрешность, мы полностью не исключаем погрешность от изменения фактора T , т. е. сама поправка a является случайной величиной:

$$a = -\theta = -(m_0 \pm z\sigma_0), \quad (2.141)$$

где σ_0 — среднее квадратическое отклонение определения поправки.

В сумме неисключенные систематические погрешности могут составлять не такую уж малую величину, заметно искажая не только результат определения математического ожидания измеряемой величины, но и, вообще говоря, его дисперсию.

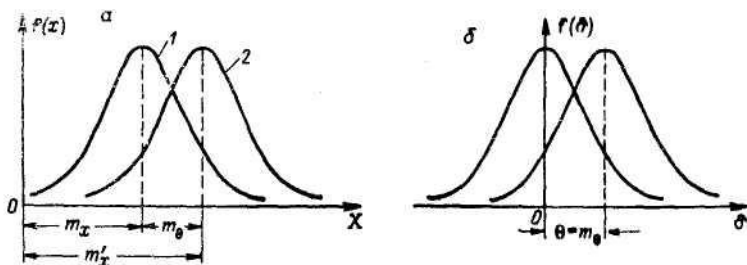


Рис. 2.19. К влиянию неисключенной остатка систематической погрешности.

Таким образом, если при $\theta = 0$ результат измерения определен как

$$X = m_X \pm z\sigma_m, \quad (2.142)$$

то при неисключенных систематических погрешностях получим

$$X' = m'_X \pm z\sigma'_m, \quad (2.143)$$

причем $m'_X - m_X = m_0$ есть математическое ожидание неисключенной систематической погрешности (рис. 2.19). Изменение σ'_m по сравнению с σ_m обычно незначительно, так как дисперсия определяется действием большого числа факторов, и будет ли их немного больше или немного меньше, не существенно, если уровень их значимости в определении дисперсии результата приблизительно одинаков или $\sigma_0 \ll \sigma_m$. Поправку на неисключенную систематическую погрешность m_0 имеет смысл вводить, если она уменьшает доверительные границы погрешности, определяемые $z\sigma_m$. Если

$$\sigma_0 \leq \sigma_m, \quad (2.144)$$

то поправку следует вводить, если они больше половины доверительной погрешности:

$$a > 0,5z \sigma_0. \quad (2.145)$$

При любом соотношении между σ_0 и σ_m следует пользоваться выражением

$$a > z\sigma_m \left(\sqrt{\frac{1 + \sigma_0^2}{\sigma_m^2}} - 1 \right). \quad (2.146)$$

Обнаружение и определение неисключенных систематических погрешностей, изменяющихся в процессе измерения, может быть сделано при обработке экспериментальных данных статистическими методами с использованием регрессионного и корреляционного анализов или метода наименьших квадратов. Неизменяющиеся в процессе измерения систематические погрешности можно обнаружить и оценить только путем измерения той же величины другим методом.

2.1.15. Запись результата измерения

Результат измерения должен быть представлен в виде математического ожидания (наиболее достоверного значения) измеряемой величины и погрешности измерения, определенной с некоторой доверительной вероятностью. При этом на величину неисключенной систематической погрешности поправку вводить не следует — необходимо ее лишь указать. Тогда запись результата измерения принимает вид

$$X = \hat{m}_X + \hat{\theta} \pm z\sigma_m \quad (P = \dots). \quad (2.147)$$

Только в том случае, когда

$$\hat{\theta} < 0,8\sigma_m, \quad (2.148)$$

можно пренебречь систематической составляющей, и, наоборот, если

$$\hat{\theta} > 0,8\sigma_m,$$

можно пренебречь случайной составляющей. Это позволит использовать полученные сведения о погрешностях измерений для оценки их при последующих измерениях тем же методом и тех же условиях.

Метод оценки погрешностей измерения, определения интервала по оценкам, полученным на основании предыдущих статистических исследований, учитывает возможный элемент случайности. Вместо z в

(2.147) подставляется так называемый толерантный множитель k (определяемый расчетным путем или по приложению 5) для доверительной вероятности P_0 того, что интервал $m_x \pm k\sigma_m$ будет содержать заданную долю генеральной совокупности P .

При выражении результатов измерения в цифровой форме следует принять погрешность округления числа, выражающего погрешность измерения, не более 15%. Это значит, что абсолютные погрешности, имеющие в старшем значащем разряде цифры 1 или 2, должны записываться двумя цифрами, например 0,12; 2,5; 0,016; погрешности, имеющие в старшем значащем разряде цифры 3, 4, 5, 6, 7, 8, 9, должны записываться одной цифрой, например: 4; 3; 0,4; 0,08 (это не относится к погрешностям, получаемым при промежуточных вычислениях).

Число, выражающее наиболее достоверное значение X , записывается таким количеством значащих цифр, которое соответствует погрешности измерения. Например, при $m_U = 101,595В$ и $z\sigma_m = 0,3В$

$$U = (101,6 \pm 0,3) В. \quad (1.149)$$

При этом погрешность округления m_x всегда будет по крайней мере на порядок меньше погрешности измерения.

Форма представления результатов измерения зависит от способа представления погрешности и определяется назначением измерений и характером использования результатов :

1. Точность определяется интервалом, в котором с заданной вероятностью находится суммарная погрешность измерения. Результат измерения представляется в виде

$$A; \Delta \text{ от } \Delta_n \text{ до } \Delta_n; P. \quad (2.150)$$

2. Точность определяется интервалом, в котором с заданной вероятностью лежит систематическая составляющая погрешности (выше указывалось, что остаточная систематическая составляющая является случайной величиной. Распределение ее принималось равномерным и учитывалось при оценке погрешности только ее математическим ожиданием m_θ), функцией распределения случайной составляющей и ее средним квадратическим отклонением σ_θ . Результат измерения представляется в виде

$$A; \theta \text{ от } \theta_n \text{ до } \theta_n; P; \\ \sigma_\theta; f(\theta). \quad (2.151)$$

3. Погрешность измерения характеризуется функциями распределения систематической и случайной составляющих погрешности и их средними квадратическими отклонениями. Результат намерения представляется в виде

$$\begin{aligned} &A; f(\delta); \sigma_{\delta}; \\ &f(\theta); \sigma_{\theta}. \end{aligned} \quad (2.152)$$

4. Точность измерения выражается функциями распределения систематической и случайной составляющих погрешности. Результат измерения записывается в виде

$$A; f(\theta); f(\delta). \quad (2.153)$$

Существующим стандартом предусмотрено использование для описания погрешностей некоторых стандартных аппроксимаций функций распределения погрешностей.

2.2. Погрешности измерительных устройств при неизменном во времени значении измеряемой величины

2.2.1. Структурные схемы измерительных устройств

Измерительное устройство (ИУ) — это устройство, обеспечивающее получение информации о значении измеряемой величины независимо от того, имеется на выходе указатель или нет, и независимо от того, куда поступает информация. Если указатель на выходе есть, то ИУ называют прибором, если нет — за ИУ либо сохраняют название «измерительное устройство», либо называют «измерительной аппаратурой» или «измерительным каналом».

Измерительные устройства для измерения как электрических, так и неэлектрических величин электрическими методами выполняются в соответствии с методами прямого и уравнивающего преобразования. На рис. 2.20 изображены структурные схемы ИУ прямого (а) и уравнивающего (б) преобразований.

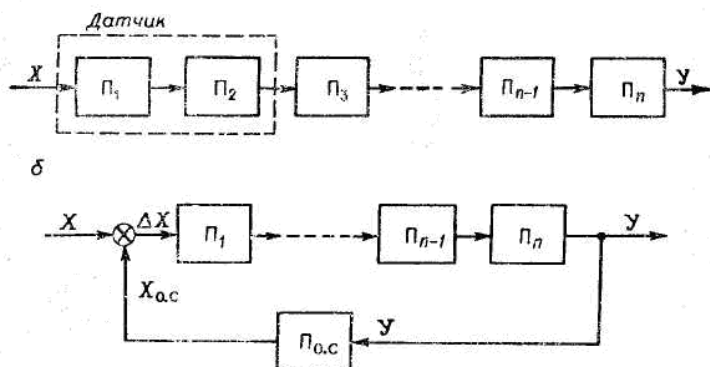


Рис. 2.20. Структурные схемы измерительных устройств. а — прямое преобразование; б — уравнивающее преобразование.

Для ИУ по схеме прямого преобразования характерно *последовательное* соединение преобразователей $\Pi_1, \Pi_2, \dots, \Pi_n$ между собой и, следовательно, последовательное прохождение измеряемого сигнала в процессе преобразования через все преобразователи канала. Если ИУ предназначено для измерения неэлектрической величины, то первые два преобразователя составляют датчик. При этом первый преобразователь играет роль чувствительного элемента, воспринимающего измеряемый физический параметр, а второй является вторичным преобразователем датчика, преобразующим выходной параметр чувствительного элемента в электрический сигнал или параметр. Чаще всего чувствительный элемент преобразует измеряемую величину в перемещение подвижной части вторичного преобразователя. Например, в датчике давления измеряемое давление преобразуется в перемещение верхнего центра мембраны, являющейся чувствительным элементом. Тип вторичного преобразователя может быть разным в зависимости от целого ряда факторов (индуктивный, емкостный и т. д.).

Назначение последующих преобразователей зависит от требуемой формы выходного сигнала. Если на выходе ИУ стоит указатель (Π_n), являющийся электромеханическим устройством, то выходной параметр датчика должен быть преобразован в величину, которая может воздействовать на это устройство и мощность которой была бы достаточна для его работы. Если измерительная информация с датчика поступает на регистрирующие устройства или в системы управления, то выходной параметр датчика преобразователями Π_3, \dots, Π_n преобразуется в ток, напряжение, частоту или число импульсов. Некоторые из этих преобразователей являются согласующими. Здесь же следует отметить, что отдельные преобразователи или группа преобразователей ИУ прямого преобразования могут быть охвачены отрицательной обратной связью, т. е. могут работать по методу уравнивающего преобразования (например, усилитель с отрицательной обратной связью). Общей структуры ИУ это не изменяет, но при исследовании точности канала измерения должно учитываться.

Структурная схема уравнивающего преобразования (рис. 2.20,б), кроме цепи прямого преобразования, имеет еще цепь обратного преобразования (обратную цепь, обратную связь). С помощью цепи обратной связи выходной параметр Y преобразуется в величину X_{0c} , имеющую тот же физический характер, что и

измеряемое воздействие X . Обратная связь ИУ отрицательная, поэтому на входе в общем случае в любой момент времени будет действовать разность

$$\Delta X = X - X_{0.c.}$$

Если ИУ выполнено как астатическая система, то процесс уравнивания (нарастания величины $X_{0.c.}$) будет происходить до тех пор, пока $X_{0.c.}$ не станет равной X (для идеальной системы), т. е. пока ΔX не станет равно нулю. Астатическая измерительная система характеризуется наличием интегрирующего преобразователя в основной цепи. Разновидностью астатической системы является следящая система (отрабатывающее устройство). Последняя характеризуется отсутствием преобразователя в обратной цепи, в связи с чем коэффициент обратной связи равен 1. Следящие системы широко используются в приборостроении для воспроизведения (повторения) входной величины.

В статических измерительных системах величина сигнала $X_{0.c.}$ на выходе преобразователя цепи обратной связи всегда меньше измеряемого сигнала. Выходной сигнал при этом оказывается пропорциональным рассогласованию $\Delta X = X - X_{0.c.}$

Рассмотренные структурные схемы построения ИУ охватывают практически все виды измерительной аппаратуры, как аналоговой, так и дискретной.

2.2.2. Характеристики преобразователей

Выше было показано, что каждое ИУ состоит из ряда преобразователей: электромеханических датчиков и указателей, усилителей, выпрямителей, формирующих устройств и т. п. Естественно, что точность и надежность измерительной аппаратуры в целом, ее рабочий диапазон, динамические характеристики определяются соответствующими характеристиками отдельных преобразователей и прежде всего — характеристиками электромеханических датчиков и указателей.

Электромеханические датчики и указатели обычно состоят из неподвижной и подвижной частей. На подвижную часть воздействует сила (момент), являющаяся результатом воздействия измеряемой величины. На рис. 2.21 изображена обобщенная схема устройства датчиков и указателей.

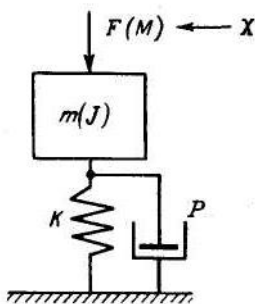


Рис. 2.21. Обобщенная схема конструкций датчиков и указателей.

Масса m (момент инерции J для подвижных частей, закрепленных на оси) через пружину, характеризующуюся жесткостью K , связана с неподвижной частью устройства. При воздействии силы F (момента M), вызванной появлением измеряемой величины, противодействующая сила (момент) пружины стремится уравновесить силу F . Вследствие инерции процесс уравновешивания сопровождается колебаниями подвижной части. Чтобы нормализовать эти колебания, т. е. обеспечить требуемый переходный процесс, в датчиках и указателях предусматриваются успокоители. На рис. 2.21 P — коэффициент успокоения.

Движение подвижной части может быть представлено дифференциальным уравнением

$$m \frac{d^2 y(t)}{dt^2} + P \frac{dy(t)}{dt} + Ky(t) = B_1 x(t) \quad (2.154)$$

(для поступательного движения подвижной части) и

$$J \frac{d^2 \alpha(t)}{dt^2} + P \frac{d\alpha(t)}{dt} + K\alpha(t) = B_2 x(t) \quad (2.155)$$

(для подвижной части, закрепленной на оси). B_1 и B_2 — коэффициенты в выражениях для силы и момента.

Введем общепринятые обозначения:

круговая частота собственных колебаний :

$$\omega_0 = \sqrt{\frac{K}{m}} \text{ или } \omega_0 = \sqrt{\frac{K}{J}}$$

степень успокоения:

$$\beta = \frac{P}{2\sqrt{KJ}}$$

или

$$\beta = \frac{P}{2\sqrt{Km}}$$

чувствительность (коэффициент преобразования) преобразователя при неизменном значении измеряемой величины:

$$S_x(0) = \frac{B_1}{K} \text{ или } S_x(0) = \frac{B_2}{K}$$

Тогда

$$\frac{d^2y(t)}{dt^2} + 2\beta\omega_0 \frac{dy(t)}{dt} + \omega_0^2 y(t) = \omega_0^2 S_x(0) x(t) \quad (2.156)$$

или в операторной форме

$$y(t)(p^2 + 2\beta\omega_0 p + \omega_0^2) = \omega_0^2 S_x(0) x(t). \quad (2.157)$$

В дальнейшем независимо от способа крепления подвижной части и физической природы выходной параметр будем обозначать $y(t)$. Передаточная функция преобразователя может быть записана в виде

$$W(p) = \frac{\omega_0^2 S_x(0)}{p^2 + 2\beta\omega_0 p + \omega_0^2}. \quad (2.158)$$

Зная передаточные функции отдельных преобразователей и вид функциональной схемы, можно найти передаточную функцию ИУ. Очевидно, что вид передаточной функции ИУ в общем случае может значительно отличаться от передаточной функции отдельного преобразователя. Дифференциальное уравнение ИУ в общем случае может быть и более высокого порядка. Динамические свойства преобразователей и ИУ будут рассмотрены ниже. Здесь же рассмотрим статические характеристики преобразователей, т. е. будем полагать, что измеряемый сигнал $x(t)$ принимает некоторое постоянное значение X :

$$x(t) = X = \text{const}, \quad (2.159)$$

а переходные процессы, характерные для реакции преобразователя на измеряемый сигнал, полностью прекратились:

$$y(t) = Y = \text{const}. \quad (2.160)$$

Тогда уравнение преобразователя (2.156) принимает вид

$$Y = W(0)X = S_x(0)X = k_x(0)X, \quad (2.161)$$

где $W(0) = S_x(0) = k_x(0) = Y/X$ — передаточная функция, чувствительность, или коэффициент преобразования преобразователя при постоянном входном воздействии. Уравнение (2.161) справедливо, если преобразователь линеен. Для нелинейного преобразователя коэффициент преобразования является функцией X . Чувствительность такого преобразователя следует определять как

$$S_x(0)=k_x(0)=dY/dX. \quad (2.162)$$

Коэффициент преобразования, чувствительность, — одна из важнейших характеристик преобразователя, в большой степени определяющая его качество, и прежде всего - точность преобразования.

Не следует путать чувствительность (коэффициент преобразования) с понятием порога чувствительности.

Порог чувствительности— это наименьшее значение измеряемой величины, способное вызвать появление сигнала на выходе ИУ. Иногда порог чувствительности называют разрешающей способностью прибора. Это не совсем точно. Понятие «разрешающая способность» является более общим, оно не связывается с началом шкалы прибора. Если порог чувствительности затруднительно оценить количественно, то разрешающая способность, как число достоверно различимых градаций результата измерений во всем диапазоне изменения измеряемой величин, может быть определена по известной погрешности преобразователя.

Важными характеристиками являются также диапазоны показаний и измерений преобразователей.

Диапазон показаний— это диапазон измеряемых величин от порога чувствительности до верхнего предела, установленного для данного преобразователя, который воспринимается им без искажения. Он определяется как

$$D_p = X_n/\Delta X_n \quad (2.163)$$

где X_n — верхний предел входной величины преобразователя; ΔX_n — порог чувствительности преобразователя..

Диапазон измерений – это диапазон измеряемых (преобразуемых) величин, внутри которого погрешность измерения (преобразования) подчиняется заданным нормам. Ширина диапазона измерений

$$D_p = X_n/X_l, \quad (2.164)$$

где X_l — нижнее значение измеряемой (преобразуемой) величины, при котором погрешность измерения (преобразования) не превышает заданной максимальной погрешности измерения (преобразования).

Максимальная погрешность измерения устанавливается экспериментатором.

Из статических характеристик, определяющих качество преобразователей, следует назвать еще одну — **мощность преобразователей**. Чем больше мощность преобразователя, тем больше мощность выходного сигнала, тем меньше соответственно влияние помех, шумов, дрейфа на точность измерения. С другой стороны, известно, что каждый последующий преобразователь, являясь

нагрузкой на предыдущий, может вызвать заметные искажения в работе последнего. Таким образом, мощность преобразователя определяет количественно два вида погрешностей измерительных преобразователей. Если первый источник погрешности (шумы, помехи, дрейф) может быть исключен или влияние его может быть снижено за счет рационального устройства преобразователя, его экранирования и т. п., то второй, связанный с реакцией предыдущего преобразователя на подключение последующего, может быть исключен только путем согласования преобразователей по мощности. При этом необходимо иметь в виду, что входное сопротивление последующего преобразователя должно быть много больше выходного сопротивления предыдущего преобразователя, если речь идет не о передаче необходимой мощности для работы электромеханического преобразователя. В последнем случае, естественно, названные сопротивления должны быть по возможности близки друг другу. Погрешность преобразователя, являющуюся также одной из основных характеристик прибора, рассмотрим отдельно.

2.2.3. Погрешности преобразователей. Аддитивная и мультипликативная составляющие

Погрешность преобразователей является следствием несовершенства их конструкции и технологии изготовления. Поэтому она определяется совокупностью частных составляющих погрешности или, как принято говорить, совокупностью частных погрешностей. Наличие погрешности у преобразователя (а она всегда есть) проявляется в том, что реальная характеристика преобразователя отличается от номинальной, является неоднозначной и из линии превращается в полосу неопределенности.

Частные погрешности можно классифицировать по различным признакам:

- 1) по характеру влияния на уравнение преобразователя;
- 2) по характеру проявления: систематические и случайные;
- 3) по причине возникновения;
- 4) по зависимости от скорости изменения измеряемой величины: статические и динамические.

По характеру влияния на уравнение преобразователя погрешности подразделяются на *аддитивные* и *мультипликативные*.

Аддитивная погрешность (от лат. *additio* — прибавление) проявляется в смещении нулевого или условно нулевого положения. Это смещение не зависит от значения измеряемой величины и объясняется наличием внешних помех, шумов, трения, порога чувствительности. К числу аддитивных можно отнести и погрешность

дискретности (квантования), хотя это и не погрешность нуля. С учетом аддитивной погрешности уравнение (2.161) преобразователя принимает вид

$$Y = S_n X + \Delta_{y.a.} \quad (2.165)$$

где $\Delta_{y.a.}$ — аддитивная погрешность, приведенная к выходу.

Аддитивная погрешность может иметь как систематический, так и случайный характер. На рис. 2.22,а показаны номинальная и реальная характеристики преобразователя для случая систематической аддитивной погрешности, а на рис. 2.22,б — полоса неопределенности, в которую превращается номинальная характеристика преобразователя, если аддитивная погрешность носит случайный характер.

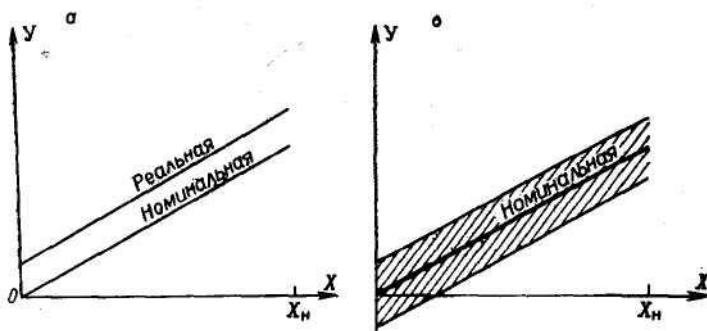


Рис. 2.22. Характеристики преобразователем при наличии аддитивной погрешности систематического (а) и случайного (б) характеров.

Систематическая составляющая аддитивной погрешности должна быть скорректирована перед началом измерения, а случайная может быть учтена по законам случайных ошибок. Перечисленные выше аддитивные погрешности являются случайными с отличным от нуля математическим ожиданием.

Мультипликативная погрешность — это погрешность чувствительности (от англ. multiplier — множитель, коэффициент), т. е. это погрешность, вызванная непостоянством чувствительности в диапазоне измерения вследствие несовершенства технологии изготовления преобразователя, а также вследствие воздействия внешних факторов.

Если непостоянство чувствительности по шкале обозначить через ΔS , то относительное изменение ее (по отношению к номинальному значению чувствительности S_n , ее математическому ожиданию) и

является относительной мультипликативной погрешностью. Действительно,

$$Y = (S_H \pm \Delta S) X = S_H X + \Delta S X = m_y + \Delta_{y,m}, \quad (2.166)$$

где $m_y = Y_0$ — математическое ожидание Y , его действительное значение; $\Delta_{y,m}$ — абсолютная погрешность преобразования.

Относительная погрешность преобразователя

$$\gamma_M = \frac{\Delta_{y,m}}{Y_0} = \frac{\Delta S}{S_H}, \quad (2.167)$$

т. е. равна относительному изменению чувствительности. Из (2.166) следует, что абсолютная мультипликативная погрешность пропорциональна измеряемой величине:

$$\Delta_{y,m} = \frac{\Delta S X S_H}{S_H} = \gamma_M S_H X = \gamma_M Y_0. \quad (2.168)$$

Здесь и ранее $\Delta_{y,m}$, $\Delta_{y,a}$, $\gamma_{y,a}$, $\gamma_{y,m}$ — это погрешности преобразователя, приведенные к выходу. Погрешности, приведенные к входу, в S_H раз меньше.

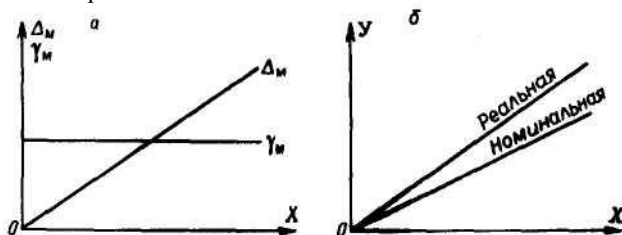


Рис. 2.23. Мультипликативные систематические погрешности (а) и характеристики преобразователей (б).

Мультипликативная погрешность также может иметь систематическую и случайную составляющие. На рис. 2.23, а изображены кривые абсолютной и относительной систематической мультипликативной погрешностей для $\gamma_M = \text{const}$, а на рис. 2.23, б номинальная и реальная характеристики преобразователя для γ_M . Если непостоянство чувствительности по шкале носит случайный характер, как это показано на рис. 2.24, а, и характеризуется среднеквадратичным отклонением $\pm \sigma_M$, то

$$\Delta_{y,m} = \pm z \sigma_M Y_0. \quad (2.169)$$

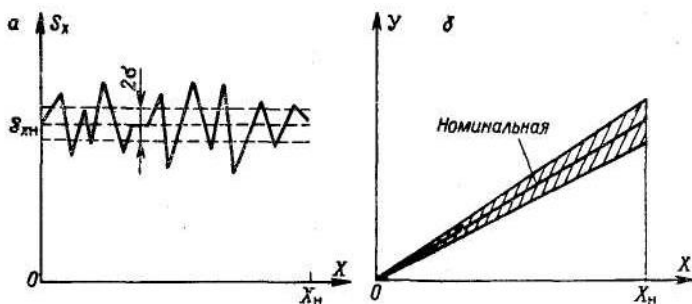


Рис. 2.24. Чувствительность (а) и характеристика преобразователя (б) при случайной мультипликативной погрешности.

На рис. 2.24,б изображена номинальная характеристика преобразователя и зона неопределенности, определяющая положение (случайное) реальной характеристики.

Полная абсолютная погрешность преобразователя, приведенная к выходу,

$$\Delta_y = \Delta_{y,a} + \gamma_m Y_0. \quad (2.170)$$

а приведенная к входу

$$\Delta_x = \Delta_{x,a} + \gamma_m X. \quad (2.171)$$

Относительная погрешность преобразователя

$$\gamma = \frac{\Delta_{y,a}}{Y} + \gamma_m = \frac{\Delta_{x,a}}{X} + \gamma_m. \quad (2.172)$$

В дальнейшем индексы y и x у погрешностей будем опускать.

Из (2.172) видно, что при малых значениях измеряемой величины относительная аддитивная составляющая погрешности может принимать очень большие значения. На рис. 2.25 изображены номинальная характеристика и полоса неопределенности, определяющая реальную характеристику, при наличии у преобразователя обеих составляющих погрешности.

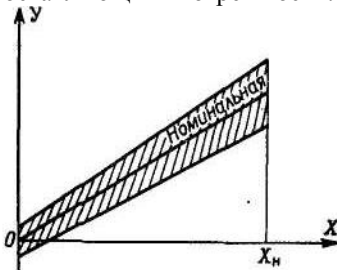


Рис. 2.25. Номинальная характеристика и полоса неопределенности реальной характеристики преобразователя при наличии аддитивной и мультипликативной погрешностей.

Погрешность, вызванная нелинейностью, возникает в том случае, когда за характеристику преобразователя, имеющего принципиально нелинейную характеристику, принимается линейная. В зависимости от способа линейризации эта погрешность может иметь только мультипликативную или только аддитивную составляющие. Действительно, при линейризации по касательной (рис. 2. 26, а) и по хорде (рис. 2.26,б) ошибка должна расцениваться как мультипликативная, имеющая систематический характер. При линейризации, например, по методу Чебышева погрешность является аддитивной (рис. 2.26, в).

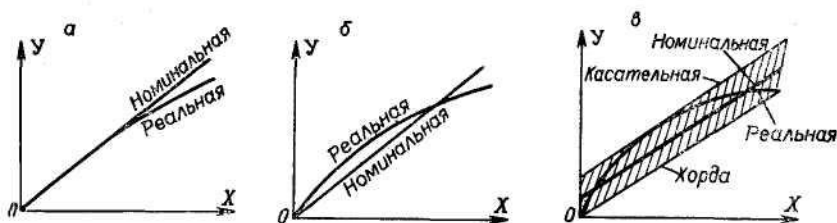


Рис. 2.26. Влияние способа аппроксимации нелинейной характеристики на характер и величину погрешности.
(Пояснения в тексте).

В этом случае она характеризуется зоной, определяемой положениями касательной и хорды, поэтому удобнее и правильнее считать частную погрешность от нелинейности при таком способе линейризации случайной величиной.

Для многих преобразователей характерно явление гистерезиса, вызывающее вариацию значений выходного параметра. Это — упругий гистерезис мембран, магнитный гистерезис ферромагнитных материалов и т. д. Замена реальной гистерезисной характеристики идеальной приводит к случайной мультипликативной ошибке.

Разделение погрешностей на мультипликативные и аддитивные очень существенно при решении вопроса о нормировании погрешностей измерительных устройств, о выборе метода оптимальной обработки получаемой информации о значении измеряемой величины.

2.2.4. Частные погрешности. Причины их возникновения и законы распределения

По причине возникновения частные погрешности можно разделить на несколько групп:

- 1) погрешности, вызванные несовершенством конструкции и технологии изготовления ;
- 2) погрешность градуировки (при индивидуальной градуировке) или шкаловая (при массовом производстве);
- 3) методические погрешности;
- 4) погрешности, вызванные старением элементов;
- 5) погрешности, вызванные влиянием внешних факторов;
- 6) погрешности отсчета.

Частные погрешности преобразователей носят случайный характер, т. е. величина их меняется случайным образом вне зависимости от значения измеряемой величины и условий работы. В большинстве случаев систематическую составляющую частной погрешности удастся исключить путем корректировки в процессе изготовления измерительного устройства или перед его использованием. Исключение составляют такие аддитивные погрешности, как порог чувствительности, шумы, трение. Полностью исключить такого рода систематические погрешности на всем диапазоне измеряемой величины невозможно. На систематические погрешности, вызванные влиянием внешних факторов, могут быть внесены поправки, если изменение влияющего фактора может быть определено. Однако делается это, как правило, только при высокоточных измерениях в лабораторных условиях. Во всех других случаях поправка на погрешность, вызванную изменением условий работы, не вводится либо потому, что нет возможности производить необходимые расчеты, либо потому, что нет возможности следить за изменением фактора, особенно при его быстрых изменениях. Тогда систематическую погрешность переводят в разряд случайных и включают в полную погрешность прибора при нормировании его точности.

Случайные погрешности, наоборот, часто проявляют себя как систематические, вызывая нелинейность реальной характеристики прибора. Это погрешности, связанные чаще всего с технологическим разбросом значений размеров деталей, размеров, характеризующих их взаимное положение, с подгонкой мер под заданный номинал и т. д. Все технологические разбросы носят случайный характер и задаются обычно величиной среднего разброса, равной половине поля допуска. При установке же, например,

конкретного резистора в схему он вызовет систематическую погрешность, связанную с отличием его действительного значения от номинального на вполне определенную величину. Конечно, в случае резистора ошибка может быть легко устранена изменением регулировочного (подгоночного) сопротивления. В других случаях это не всегда удается или удастся не в полной мере. Например, не удастся полностью устранить эксцентриситет между центром расточки полюсных наконечников магнитной системы и осью цилиндрического сердечника, являющийся случайной величиной (рис. 2.27,а).

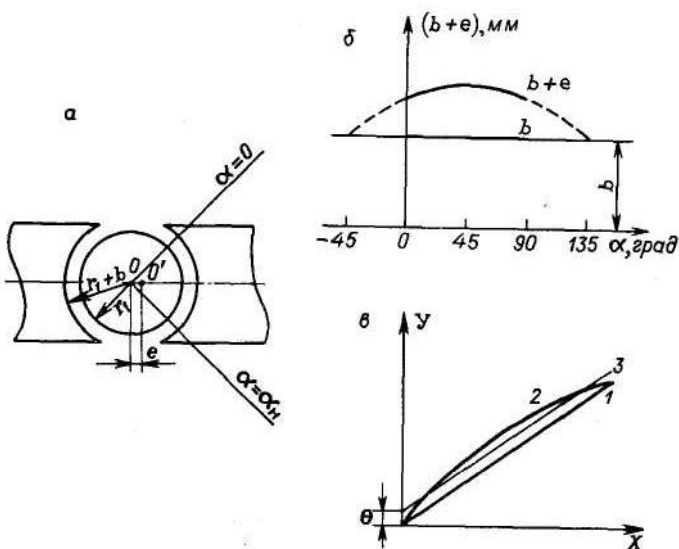


Рис. 2.27. К влиянию эксцентриситета на характеристику преобразователя. (Пояснения в тексте).

Это приводит к тому, что расстояние между поверхностями сердечника и наконечников будет определяться постоянной составляющей b и переменной, равной $e \sin 2\pi\alpha/\alpha_n$, где α — угловая координата; α_n — диапазон изменения координаты α (рис. 2.27,б). Соответственно изменяется индукция на рабочем участке, определяющая чувствительность S_x , в связи с чем реальная характеристика принимает вид кривой 2 на рис. 2.27, в. При аппроксимации кривой 2 по методу Чебышева получается кривая 3, сдвинутая относительно номинальной на величину θ на которую путем регулировки вводится поправка.

Остается составляющая, связанная с отклонениями реальной кривой 2 от кривой 3 или 1, так как путем введения поправки $a = -\theta$ они совмещаются. Ее считают случайной погрешностью.

Такой подход к анализу частных погрешностей сложен, но соответствует действительным преобразованиям ошибок и потому является единственно правильным. Естественно, при этом встает вопрос — как выбрать закон распределения полученной частной погрешности, зависит ли он от закона распределения влияющего фактора или от характера изменения его в диапазоне преобразования.

Определение законов распределения, как уже указывалось, производится на основании статистических исследований. Однако поставить опыт так, чтобы погрешность измерения определилась только изучаемым фактором, очень трудно. Это связано с значительным усложнением методики опыта и обработкой результатов измерений. Очень часто определить экспериментально закон распределения той или иной частной погрешностям вообще не представляется возможным. В этих случаях приходится принимать гипотетические законы распределения, основываясь на изучении фактора, вызывающего погрешность, характера его влияния на характеристику преобразователя (прибора). Следует заметить, что один и тот же источники ошибки может вызвать в разных приборах (преобразователях) частные погрешности, имеющие различные законы распределения, т. е. преобразователи по-разному трансформируют воздействия.

Законы распределением факторов (внешних и внутренних, т. е. технологических) могут быть различными. Законы распределением технологических факторов и погрешностей, вызываемых ими, достаточно хорошо изучены. Законы же распределения таких факторов, как температура, колебания напряжения источников питания и т. п. определяются условиями работы аппаратуры и, естественно могут принимать различные формы. Поэтому законы распределения их могут быть определены только на основании опытных данных, полученных для конкретных условий работы. Если статистика достаточно велика, то обычным путем строится гистограмма распределения влияющего фактора. Затем подбирается простейший теоретический закон, наиболее хорошо совпадающий с полученной гистограммой. Для нормального закона все может быть сделано с достаточной строгостью, как это было изложено в разделе 1; установление законов распределения отличных от нормального, является приближенным. Так, распределение напряжения сети 220 В,

50 Гц в течение суток имеет форму, изображенную на рис. 2.28 сплошной линией.

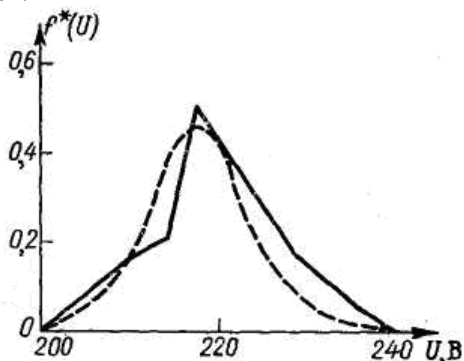


Рис. 2.28. Статистический закон распределения напряжения сети.

Ни один теоретический закон не подходит для его описания. Для приближенных оценок можно принять закон распределения напряжений нормальным, хотя проверка гипотезы правдоподобия и даст отрицательный результат.

Если внешние факторы изменяются по детерминированным законам, но погрешности, вызванные ими, учитываются как случайные (поправки не вводятся), то необходимо определить законы распределения последних для заданных детерминированных воздействий. Допустим, что температура за время работы преобразователя изменяется линейно (рис. 2.29, а) от T_{\min} до T_{\max} , причем при $T=T_{\text{тип}}$ температурная погрешность Δ_T равна нулю. Тогда при изменении температуры Δ_T будет линейно возрастать.

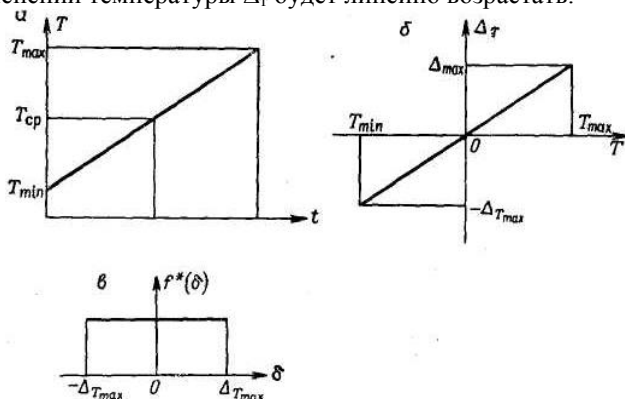


Рис. 2.29. Установление закона распределения температуры окружающей среды.

Для перевода Δ_T в разряд случайных погрешностей (хотя бы по знаку) при проектировании прибора делается коррекция на величину температурной погрешности, соответствующей изменению температуры на

$$T_{\max} - \frac{T_{\max} - T_{\min}}{2} = T_{\max} - T_{\text{ср}}.$$

Тогда при $T=T_{\text{ср}}$ температурная погрешность будет равна нулю (рис. 2.29,б); при $T < T_{\text{ср}} \Delta_T = \delta_T < 0$, при $T > T_{\text{ср}} \Delta_T = \delta_T > 0$, причем максимальные абсолютные значения погрешностей не превысят значений, соответствующих изменению температуры на $\Delta_T = |T_{\max} - T_{\text{ср}}|$. Вероятность появления разных по величине погрешностей, соответствующих разным значениям температуры, как линейной функции времени, одинакова, т. е. погрешности распределены по закону равномерной плотности. При случайном изменении температуры относительно $T_{\text{ср}}$ закон распределения температурной погрешности будет другим, нам не известным. Чтобы иметь возможность учесть температурную погрешность и в этом случае, задают предельные значения δ_T , соответствующие T_{\max} и T_{\min} , а распределение температур в диапазоне изменения принимают также равновероятным. При этом предельное значение погрешности равно $\delta_{T_{\max}}$ (рис. 2.29, в), среднее квадратическое отклонение $\sigma = \Delta_{T_{\max}}/\sqrt{3}$. Аналогично рассуждая, можно найти приближенные законы распределения и для других частных погрешностей, вызванных внешними и внутренними факторами.

Распределим частные погрешности по принадлежности к законам распределения.

I. Нормальный закон распределения

1. Погрешность прибора, складывающаяся из большого числа частных составляющих при отсутствии доминирующей частной погрешности.

2. Погрешности, вызванные флюктуациями эмиссии электродов ламп, флюктуациями за счет дробового эффекта, тепловыми шумами. Это — мультипликативные погрешности, задаваемые обычно предельным отклонением (допуском) $\pm \delta_{\max}$. Среднее квадратическое отклонение в поле допуска $\sigma = \delta_{\max}/3$ при $P = 0,997$.

II. Закон равномерной плотности распределения

1. Погрешность подгонки сопротивлений под заданный номинал. Мультипликативная: задается допуском $\pm \delta_{\max}$, $\sigma = \delta_{\max}/\sqrt{3}$.

2. Погрешность отсчета по шкале долей деления и на цифровых отметках. Аддитивная. Задается величиной $\pm \delta_{\max} = 0,1 \div 0,5$ деления в зависимости от качества отсчетного устройства; $\sigma = \delta_{\max} / \sqrt{3}$.

3. Порог чувствительности, погрешность уравнивания. Аддитивные; задаются обычно минимальным значением напряжения U (угла отклонения α), при котором следующий преобразователь начинает устойчиво работать (наблюдатель — уверенно отсчитывать). Максимальное из минимальных значений U или α принимается за $\pm \delta_{\max}$ (рис. 2.30). Внутри допуска $\sigma = \delta_{\max} / \sqrt{3}$.

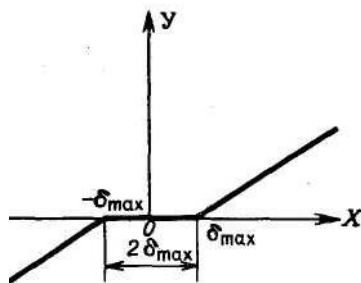


Рис. 2.30. Характеристика преобразователя с учетом порога чувствительности.

4. Погрешность дискретности, витковая погрешность потенциометров. Аддитивная. Первая связана с квантованием измеряемой величины по уровню (см. рис. 2.31, *a*) и определяется величиной младшего кванта (единицей последнего разряда); вторая вызвана дискретным (а не линейным) изменением сопротивления при перемещении α щетки с витка на виток (см. рис. 2.31, *б*). Если величину кванта и сопротивление витка обозначить через $2\delta_{\max}$, то погрешность дискретности и витковая не превзойдут величины $\pm \delta_{\max}$; в диапазоне $2\delta_{\max}$ погрешности разных значений равновероятны, среднее квадратическое отклонение $\sigma = \delta_{\max} / \sqrt{3}$.

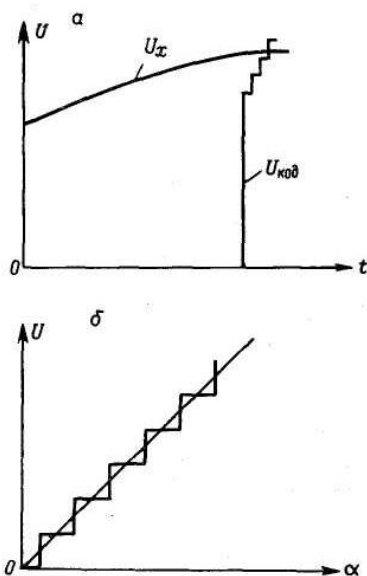


Рис. 2.31. Погрешности квантования (а) и витковая (б).

5. Погрешность от трения. Аддитивная. Характеризуется и задается величиной зоны застоя $2\delta_{\max}$ («недохода»), образующейся, если момент (сила) трения оказывается больше устанавливающего момента (силы). На рис. 2.32 прямая 1 соответствует преобразователю без трения (идеальному); если принять, что трение случайно, но постоянно в диапазоне преобразования X , то характеристика реального преобразователя будет начинаться в одной из случайных точек зоны застоя $O'O''$ и пойдет параллельно прямой 1.

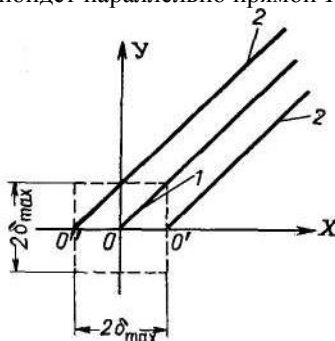


Рис. 2.31. Влияние зоны застоя на характеристику преобразователя.

Прямые 2 соответствуют предельной величине зоны застоя, т. е. $M_{тр, max}$. Внутри зоны застоя погрешности от трения распределены равномерно; $\sigma = \delta_{max} / \sqrt{3}$. На самом деле погрешность от трения не остается постоянной в диапазоне изменения измеряемой величины и зависит от условий работы преобразователя, что приводит к очень приближенным оценкам погрешности от трения.

6. Погрешность люфта кинематической пары. Аддитивная.

Обычно принимают, что вследствие люфта характеристика преобразователя принимает вид, изображенный на рис. 2.33, а кривой 2. Прямая 1 — характеристика преобразователя без люфта. При этом вероятности появления погрешностей, равных $+\delta_{max}$ и $-\delta_{max}$, будут составлять 0,5, а вероятности появления погрешностей, меньших $|\delta_{max}|$, равны нулю, т. е. закон распределения погрешностей от люфта может быть представлен так, как это показано на рис. 2.33, б. Среднее квадратическое отклонение $\sigma = \pm \delta_{max}$. Закон равномерной плотности вырождается в двухмодальный, дискретный.

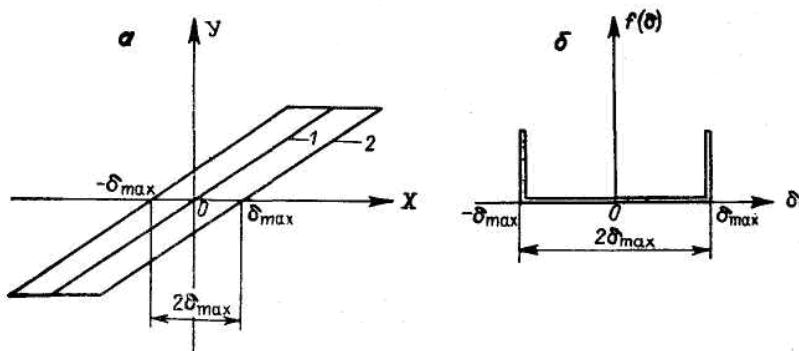


Рис. 2.33. Искажение характеристики при наличии люфта (а) и закон распределения погрешностей от люфта (б).

7. Погрешность от гистерезиса (магнитного, упругого последовательного действия) аналогична погрешности люфта. Однако так как гистерезисные кривые нелинейны, то закон распределения погрешностей, оставаясь двухмодальным, принимает несколько иную форму (рис. 2.34).

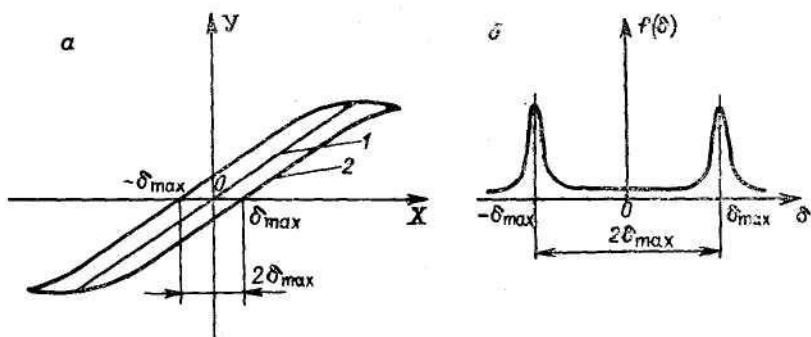


Рис. 2.34. Влияние гистерезиса на характеристику преобразователя (а) и закон распределения погрешности от гистерезиса (б).

Аддитивная; предельное значение равно $\pm \delta_{\max}$ (задается); среднее квадратическое отклонение может быть подсчитано как $\sigma = \delta_{\max}$, если принять двухмодальный дискретный закон (см. п. б), или как $\sigma = \delta_{\max}/\sqrt{3}$, если закон распределения принять равномерным. Выбор закона зависит от степени нелинейности из-за гистерезиса реальной характеристики преобразователя.

8. Неисключенные остатки систематических погрешностей. Аддитивные. Описание их было сделано в 2.1.13.

9. Погрешности, вызываемые колебаниями температуры внутри заданных пределов, когда закон распределения значений температуры не задан (см. выше).

10. Погрешности, вызываемые колебаниями напряжения внутри заданных пределов, когда закон распределения значений напряжения не задан (аналогично п. 9).

11. Погрешности, вызываемые колебаниями любого фактора внутри заданных пределов, если закон распределения его неизвестен. В частности, неизвестным является закон распределения погрешности градуировки (индивидуальной), складывающейся из ряда составляющих: погрешности образцового прибора, погрешности вычерчивания шкалы (равномерности делений), погрешности установки. Эти составляющие имеют разные законы распределения. Если принять, что доминирующей погрешностью является погрешность установки шкалы, имеющая равномерный закон распределения, то и погрешность градуировки следует принять распределенной по тому же закону. Если допустить, что все составляющие погрешности имеют один вес и распределены по закону

равномерной плотности, то закон распределения будет близким к нормальному.

III. Закон распределения Симпсона

Это композиция двух законов равномерной плотности с одинаковыми предельными значениями ошибок $\pm \delta_{\max}$

1. Погрешности измерения величин по двум отсчетам. Например, длины угла, отрезка времени. Это погрешность выполнения аппроксимации графоаналитическим методом и др. Предельная погрешность равна $\pm 2\delta_{\max}$ (рис. 2.35). Среднее квадратическое отклонение $\sigma = 2\delta_{\max} / \sqrt{6}$

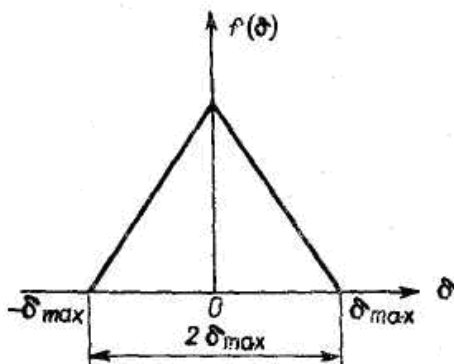


Рис. 2.35. Закон распределения Симпсона.

2. Погрешности измерения методом замещения, когда измеряемая величина определяется по двум отсчетам меры.

IV. Закон распределения арксинуса (арккосинуса)

Плотность распределения закона арксинуса определяется уравнением

$$f(\delta) = \frac{1}{\pi \sqrt{1 - \delta^2}}. \quad (2.173)$$

Кривая, соответствующая этому закону, приведена на рис. 2.36. Если предельное значение погрешности $\pm \delta_{\max}$, то $\sigma = \delta_{\max} / \sqrt{2}$

По закону арксинуса распределены все погрешности, величина которых меняется по синусоидальному закону. По отношению к моменту измерения фаза таких погрешностей является случайной величиной. Поэтому случайной является и величина погрешности в момент измерения. Вероятность того, что в момент измерения погрешность равна нулю, мала, так как синусоидальная величина меняет знак достаточно быстро. Наоборот, в области максимума кривая имеет

пологий участок, т. е. скорость изменения величины погрешности здесь мала и вероятность появления больших погрешностей больше.

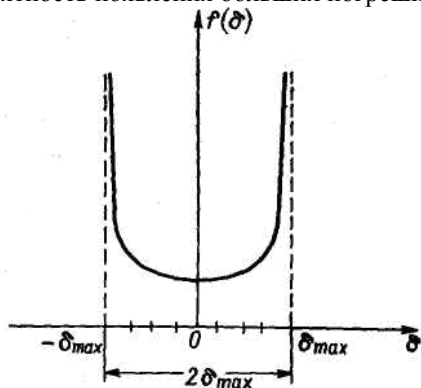


Рис. 2.36. Закон распределения арксинуса.

В качестве примера можно назвать аддитивную погрешность, вызванную синусоидальной наводной (помехой) $u = U_m \sin \omega t$. Принимая плотность распределения обратно пропорциональной скорости изменения напряжения, получим :

$$f(U) = \frac{A}{dU/dt} = \frac{A}{U_m \cos \omega t} = \frac{A}{U_m \cos (\arcsin u/U_m)} = \frac{A}{U_m \sqrt{1 - (u/U_m)^2}}. \quad (2.174)$$

Из условия нормирования функции $f(U)$

$$\frac{A}{U_m} \int_{-U_m}^{+U_m} \frac{du}{\sqrt{1 - (u/U_m)^2}} = A\pi = 1,$$

находим $A = 1/\pi$. Тогда

$$f(U) = \frac{1}{\pi U_m \sqrt{1 - (u/U_m)^2}}. \quad (2.175)$$

Дисперсия (второй центральный момент)

$$D_u = \int_{-U_m}^{+U_m} u^2 \frac{1}{\pi U_m \sqrt{1 - (u/U_m)^2}} du = \frac{1}{2} U_m^2 \quad (2.176)$$

и

$$\sigma_u = U_m / \sqrt{2}. \quad (2.177)$$

2.2.5. Методические погрешности

Методические погрешности — это тоже частные погрешности преобразователей, приборов. Особенность их заключается в том, что по своей природе это систематические погрешности. При проектировании приборов методические погрешности стараются свести к ничтожно малой величине, однако полностью исключить их не удается. Встает вопрос об учете их величины при оценке точностных свойств приборов. Методические погрешности вызываются различными причинами. Здесь может встретиться несколько случаев.

1. Зависимость измеряемой величины от определяющих ее параметров не является абсолютно точной или преобразователь не реализует заданной зависимости измеряемой величины от определяющих параметров. Если $x = \varphi(x_1, x_2, \dots, x_n, a_1, a_2, \dots, a_n)$ есть точная зависимость x от параметров, а $x_I = \varphi_I(x_1, x_2, \dots, x_n, a_1, a_2, \dots, a_n)$ — реализуемая преобразователем, то методическая погрешность

$$\Delta_{M1} = \varphi_I(x_1, x_2, \dots, x_n, a_1, a_2, \dots, a_n) - \varphi(x_1, x_2, \dots, x_n, a_1, a_2, \dots, a_n). \quad (2.178)$$

Погрешность Δ_{M1} проявляется в отличии реальной характеристики преобразователя $y_I = \varphi_I(x_1, x_2, \dots)$ от точной зависимости $y = \varphi(x_1, x_2, \dots)$ и может быть учтена тем же способом, что и погрешность нелинейности преобразователя.

Следует напомнить, что погрешность от нелинейности, которая рассматривалась в предыдущих параграфах, была вызвана искажениями характеристики преобразователя за счет несовершенства конструкции и технологии его изготовления, но сам преобразователь по принципу действия мог воспроизводить заданную зависимость точно. Здесь же имеется в виду, что преобразователь выбран так, что он принципиально не может воспроизводить точно заданной зависимости. Например, ваттметр, предназначенный для измерения мощности, теряемой в нагрузке, не реализует заданной зависимости, так как всегда измеряет и мощность, потребляемую одной из своих обмоток. Действительно, ваттметр, включенный в схему, изображенную на рис. 2.37, a , измеряет мощность

$$\begin{aligned} Y &= S_P P = S_P I U = S_P I (I R_W + R_{нар}) = \\ &= S_P (P_W + P_{нар}), \end{aligned} \quad (2.179)$$

где R_w — сопротивление последовательной обмотки ваттметра; P_w , $P_{\text{наг}}$ — мощности, теряемые в последовательной обмотке ваттметра и нагрузке; S_p — коэффициент преобразования ваттметра.

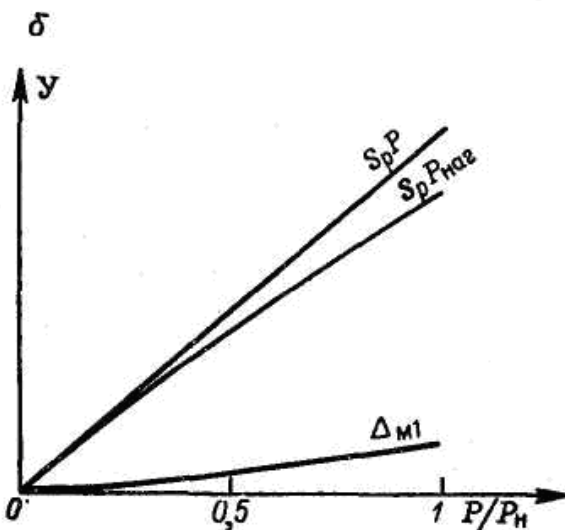
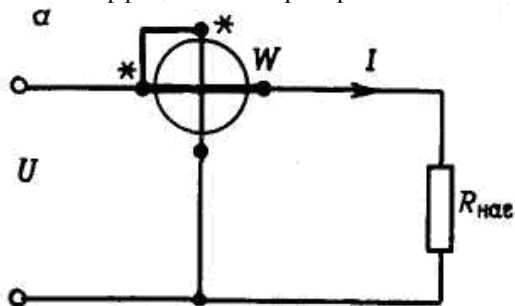


Рис. 2.37. Неточность воспроизведения заданной зависимости преобразователем. (Пояснения о тексте).

Методическая погрешность (рис. 2.37, б)

$$\begin{aligned} \Delta_{M_1} &= S_p P_W = S_p I^2 R_W \approx \\ &\approx S_p U^2 \frac{R_W}{R_{\text{наг}}^2} \quad (2.27) \end{aligned} \quad (2.180)$$

растет с уменьшением $R_{\text{наг}}$, т. е. с увеличением тока нагрузки.

2. Преобразователь точно реализует заданную зависимость x от параметров, но сами параметры вводятся в преобразователь с погрешностями, т. е. вместо x_1, x_2, \dots, x_n вводятся $x_{11}, x_{21}, \dots, x_{n1}$. Тогда

$$\begin{aligned} \Delta_{M_2} &= \varphi(x_{11}, x_{21}, \dots, x_{n1}, a_1, a_2, \dots, a_n) - \\ &- \varphi(x_1, x_2, \dots, x_n, a_1, a_2, \dots, a_n). \end{aligned} \quad (2.181)$$

Если разности $\Delta_1 = x_{11} - x_1$; $\Delta_2 = x_{21} - x_2$; ...; $\Delta_n = x_{n1} - x_n$, малы, то для определения погрешности $\varphi(x_{11}, x_{21}, \dots, x_{n1}, a_1, a_2, \dots, a_n)$ можно разложить в ряд Тейлора по степеням Δ . Тогда, ограничиваясь первыми двумя членами разложения найдем

$$\Delta_{M_2} = \frac{\partial \varphi}{\partial x_1} \Delta_1 + \frac{\partial \varphi}{\partial x_2} \Delta_2 + \dots + \frac{\partial \varphi}{\partial x_n} \Delta_n, \quad (2.182)$$

причем производные определяются при номинальных значениях определяющих параметров. Частные составляющие методической погрешности Δ_{M_2} можно принять случайными величинами, имеющими нулевые математические ожидания, если $\Delta_1, \Delta_2, \dots, \Delta_n$ связаны с неточностью измерения соответственно параметров x_1, x_2, \dots, x_n . Но бывает и так, что математические ожидания $\Delta_1, \Delta_2, \dots, \Delta_n$ не равны нулю и, наоборот, даже много больше случайных составляющих. Если $m_{\Delta_i} > 0,8\sigma_m$ (см. 2.1.14), то пренебрегать систематической погрешностью $\theta_i = m_{\Delta_i}$ нельзя. На нее нужно вводить поправку. Если это не делается, то погрешность измерения будет содержать систематическую составляющую, равную сумме θ_i . Так, для получения правильных показаний барометрического высотомера необходимо вводить непрерывно значения T_0 и P_0 на уровне земли. Если же T_0 и P_0 вводятся однажды при вылете самолета, то возникает систематическая погрешность, причем тем большая, чем больше время полета.

3. Преобразователь точно реализует заданную зависимость, но потребляет на себя некоторую мощность и тем искажает параметры измерительной цепи. Методическая погрешность при этом выражается в общем виде аналогично (2.181), но под $x_{11}, x_{21}, \dots, x_{n1}$ здесь следует понимать не вводимые извне параметры, определяющие x , а

непосредственно искаженные преобразователем параметры измерительной цепи.

Если прибор состоит из ряда преобразователей, как это показано на рис. 2.38, то каждый последующий преобразователь является нагрузкой для предыдущего и будет искажать его выходной сигнал.

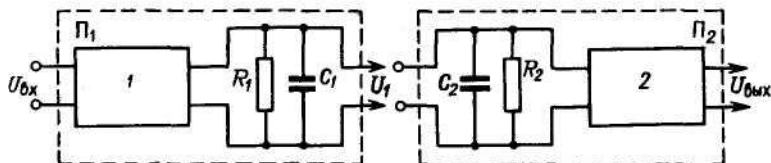


Рис. 2.38. Изменение параметров при соединении преобразователей.

Так, если выходные параметры преобразователя 1 (на рис. 2.38) — R_1 , C_1 и U_1 , то при подключении к нему преобразователя 2 с входными параметрами R_2 , C_2 выходными параметрами преобразователя 1 следует считать

$$C_{11} = C_1 + C_2; \quad R_{11} = \frac{R_1 R_2}{R_1 + R_2}.$$

Это приводит к искажению U_1 как по амплитуде, так и по фазе. В каждом конкретном случае погрешность, вызванную несогласованностью выходной цепи одного преобразователя с входной цепью следующего, можно достаточно просто подсчитать. Однако ее величина обычно не остается постоянной, так как R_1 , R_2 , C_1 , C_2 меняются в зависимости от положения движков потенциометров, от изменения режима транзисторов, и т. д. Поэтому вводить поправку затруднительно. При проектировании приборов предусматриваются обычно согласующие преобразователи (катодные и эмиттерные повторители, схемы отработки, согласующие трансформаторы), исключаящие или снижающие до ничтожно малой величины рассматриваемую методическую погрешность.

Неисключенная часть методической погрешности, как неисключенная часть систематической погрешности, принимается случайной величиной, распределенной по закону равномерной плотности.

2.2.6. Суммирование частных погрешностей

Вопрос о суммировании частных погрешностей не может быть решен однозначно независимо от того, идет ли речь об определении погрешности преобразователя или прибора, состоящего из ряда преобразователей и независимо от того, являются ли частные погрешности систематическими или случайными, зависимыми или независимыми, коррелированными или некоррелированными, аддитивными или мультипликативными. Естественно, что учесть все названные обстоятельства при определении суммарной погрешности ИУ очень трудно. Единой теории суммирования частных погрешностей не существует. Исследователи подходят к решению этого вопроса по-разному, единого мнения, несмотря на неоднократные обсуждения, не выработано.

Преобразователь — это простейшее звено прибора; его задача сводится к преобразованию входной величины X_i в X_{i+1} ; коэффициент преобразования его k_i . Если внимательно исследовать принцип действия прибора, то таких простейших звеньев можно насчитать достаточно много, причем некоторые из них не имеют специального конструктивного оформления. Например, термометр сопротивления, принципиальная схема которого изображена на рис. 2.39, состоит из четырех преобразователей: первый преобразует измеряемую температуру T_x в температуру T_R терморезистора, второй — T_R в изменение сопротивления R , третий — R в напряжение U , четвертый — U в выходной сигнал Y . Первый не имеет формы преобразователя в общепринятом смысле этого слова. Однако это преобразователь, определяющий динамику работы термометра и являющийся источником ошибок.

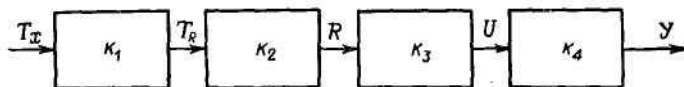


Рис. 2.39. Структурная схема термометра сопротивления.

Каждый преобразователь осуществляет преобразование входящей в него величины с некоторой мультипликативной и аддитивной случайными погрешностями, причем не обязательно с нулевым математическим ожиданием. Эти погрешности определяются частными составляющими. Схему прибора и его частные погрешности можно представить так, как это показано на рис. 2.40. При этом любое

$$\Delta = \Delta_M + \Delta_a, \text{ где } \Delta_M = \theta_M + \delta_M \text{ и } \Delta_a = \theta_a + \delta_a.$$

Общее число частных погрешностей прибора может достигать больших величин (30-50).

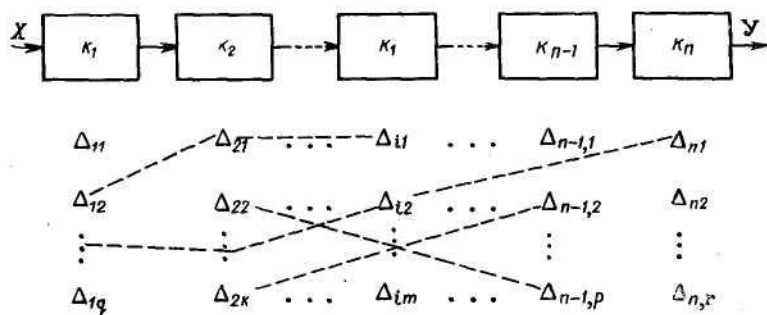


Рис. 2.40. Частные погрешности и установление корреляционных связей.

В практике измерений применяются разные способы суммирования частных погрешностей. Основные из них можно сформулировать следующим образом.

1. *Арифметическое суммирование абсолютных максимальных значений всех частных погрешностей.* Получаемое значение суммарной погрешности при этом часто называют предельной погрешностью, т. е. максимально возможной. Некоторые авторы вычисляют в дополнение среднее квадратическое значение суммарной погрешности без учета степени корреляции, характера законов распределения частных погрешностей. Таким образом, в качестве характеристик точности называются

$$\Delta = \sum_{ij} \Delta_{ij} \tag{2.183}$$

и

$$\Delta_s = \sqrt{\sum_{ij} \Delta_{ij}^2} . \tag{2.184}$$

Здесь первый индекс — номер преобразователя, второй — номер ее частной погрешности.

Если предельная погрешность \$\Delta\$ еще о чем-то свидетельствует, хотя вероятность ее появления в зависимости от числа погрешностей составляет величину порядка \$10^{-10}\$-\$10^{-20}\$, то вторая практически не несет никакой информации, так как такое геометрическое суммирование без учета названных выше обстоятельств является нестрогим, неверным.

2. Геометрическое суммирование в предположении, что все частные погрешности являются условно-случайными и распределены по нормальному закону. Тогда суммарная погрешность также имеет нормальный закон распределения и может быть представлена как

$$\Delta = z\sigma = z \sqrt{\sum_{ij} \left(\frac{\Delta_{ij}}{z_{ij}}\right)^2} = \sqrt{\sum_{ij} \Delta_{ij}^2}, \quad (2.185)$$

так как по условию $z=z_{ij}$ и выбирается в зависимости от принятой доверительной вероятности. Чаще всего принимают $z = 3$ ($P = 0,997$); получаемое при этом значение Δ_{\max} , называют максимальной погрешностью ($\Delta_{\max} = 3\sigma$).

3. Суммирование погрешностей с учетом корреляционных связей и законов распределения. Перед суммированием частных погрешностей прибора производят их анализ с точки зрения выявления законов распределения и жестких корреляционных связей, имея в виду обнаружение частных погрешностей одного или нескольких преобразователей, вызванных одним и тем же фактором, например температурой или напряжением питания. Для группы погрешностей, вызванных одним и тем же фактором и вследствие этого жестко коррелированных, принимаются один и тот же закон распределения и коэффициент взаимной корреляции, равный +1 или -1, так как вычисление действительных коэффициентов корреляции является очень трудоемким. Суммарная погрешность этой группы частных погрешностей при этом равна алгебраической сумме составляющих, т. е.

$$\Delta_{kv} = \sum_{ji} \Delta_{ijkv}, \quad (2.186)$$

где индексом k отмечены частные погрешности, относящиеся к рассматриваемой группе. Таких групп коррелированных между собой погрешностей несколько: 1, 2, 3 ..., v ... (см. рис. 2.40). Суммарные погрешности каждой группы $\Delta_{k1}, \Delta_{k2}, \dots, \Delta_{kv}, \dots$ распределены по законам, соответствующим законам определяющих их частных погрешностей, и являются практически независимыми. Не объединенные корреляционными связями частные погрешности Δ_{ijH} можно также принять статистически независимыми. Тогда среднее квадратическое отклонение суммы всех частных погрешностей прибора

$$\sigma = \sqrt{\left(\frac{\Delta_{k1}}{z_1}\right)^2 + \left(\frac{\Delta_{k2}}{z_2}\right)^2 + \dots + \sum_{ij} \left(\frac{\Delta_{ijH}}{z_{ij}}\right)^2}, \quad (2.187)$$

Если среди погрешностей, входящих в последнее выражение, нет доминирующих и законы распределения их симметричны, то для общей погрешности прибора можно принять нормальный закон распределения. Тогда, выбрав z для заданной доверительной вероятности и подсчитав σ по (2.187), получим суммарную погрешность прибора

$$\Delta = z\sigma. \quad (2.188)$$

Если среди погрешностей, определяющих σ , есть доминирующие, то для определения суммарной погрешности необходимо найти закон ее распределения как композицию законов распределения погрешностей Δ_{ijk} , Δ_{ijn} . По полученному закону можно определить σ и Δ (как максимальную погрешность). Для разных законов соотношение между σ и Δ , т. е. z , будет разным. Само определение σ и Δ при большом числе составляющих погрешностей очень трудоемко.

П. В. Новицкий, определяя погрешности измерительных устройств с позиций теории информации, установил, что коэффициент связи между суммарной погрешностью Δ и средним квадратическим отклонением σ , названный им энтропийным коэффициентом k_z , изменяется от 0 до 2,07, причем максимальное значение соответствует нормальному закону распределения (рис. 2.41). Кривая построена для плавных симметричных законов распределения и является предельной, так как энтропийные коэффициенты для несимметричных и неплавных законов лежат ниже кривой, но в области, которую она ограничивает. Энтропийные коэффициенты рассчитываются несколько проще. Следует отметить, что теория информации дает однозначное значение $z = k_z = 2,07$ для нормального закона (независимо от доверительной вероятности). Объясняется это тем, что энтропийная погрешность и, следовательно, k_z находятся как мера дезинформации, определяемая в теории информации как условная энтропия

$$H(X/X_n) = - \int_{-\infty}^{+\infty} f(\Delta) \log f(\Delta) d\Delta,$$

где X — измеряемая величина; X_n — показание прибора в единицах намеряемой величины. Таким образом, при использовании теории информации подход к определению $z = k_z$ одинаков для всех законов распределения.

4. Способ определения суммарной погрешности измерительного устройства по методу С. П. Дмитриева и Л. А. Овчинникова.

После расчета частных составляющих погрешностей преобразователей и определения их числовых характеристик σ_i и θ_i каждый реальный преобразователь заменяется идеальным с номинальным (заданным) коэффициентом преобразования k_n и

добавлением на выходе ошибки ΔY этого преобразователя так, как это показано на рис. 2.42. Погрешность ΔY может содержать аддитивную и мультипликативную составляющие.

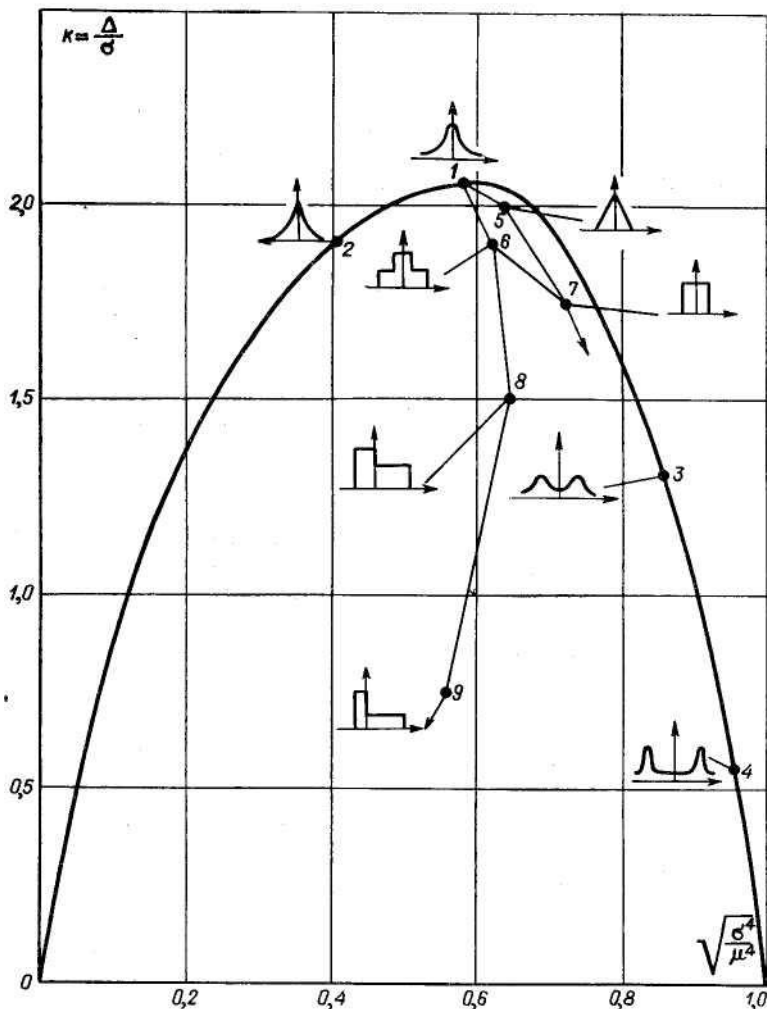


Рис. 2.41. Зависимость энтропийного коэффициента для плавных симметричных законов распределения от величины $\sqrt{\sigma^4/\mu^4}$.

Выражается она в виде математического ожидания θ_i и среднего квадратического отклонения σ_i . Для определения суммарной ошибки измерительного устройства ошибки преобразователей (каждая отдельно) трансформируются на выход прибора в соответствии с его структурной схемой, т. е. используется принцип суперпозиции, справедливый для линейных преобразователей (погрешность нелинейности учитывается как одна из частных погрешностей преобразователя).

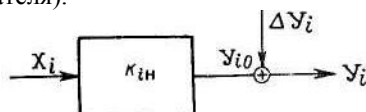


Рис. 2.42. Структура реального преобразователя.

В результате, если число преобразователей n , то погрешность прибора, приведенная к выходу, будет иметь n составляющих типа

$$\Delta Y_1 = k_2 k_3 \cdot \dots \cdot k_i \cdot \dots \cdot k_n \Delta Y_{1i};$$

$$\Delta Y_2 = k_3 k_4 \cdot \dots \cdot k_i \cdot \dots \cdot k_n \Delta Y_{2i};$$

.....

$$\Delta Y_n = k_n \Delta Y_{ni},$$

где k_i — номинальные коэффициенты преобразования (рис. 2.43).

Ошибки $\Delta Y_1, \Delta Y_2, \dots, \Delta Y_n$ принимаются независимыми, поэтому среднее квадратическое отклонение ошибки на выходе определяется как геометрическая сумма независимых величин, а систематическая — как алгебраическая сумма. При необходимости можно отдельно подсчитать σ и θ по аддитивной и мультипликативной составляющим.

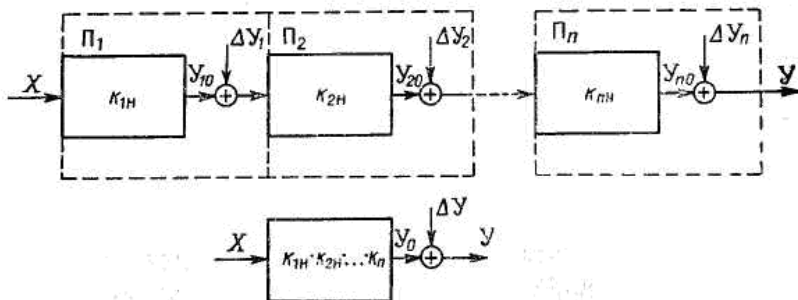


Рис. 2.43. Формирование ошибки измерительного устройства.

Очевидно, что наибольший интерес представляют третий и четвертый методы определения суммарной ошибки прибора. При этом последний больше соответствует физической стороне происходящих явлений: ошибка предыдущего преобразователя вместе с выходным сигналом поступает на вход следующего преобразователя, и так далее, вплоть до выхода прибора. Одним из недостатков такой методики является трудность установления закона распределения суммарной погрешности преобразователя, так как обычно число частных погрешностей преобразователя невелико и композиция законов распределения (распределения суммарной погрешности преобразователя) отличается от нормального закона; здесь более явно сказываются доминирующие погрешности. Вторым недостатком является невозможность учета корреляционных связей между отдельными частными погрешностями преобразователей, вызванные, например, изменениями температуры, напряжения.

Третий метод суммирования частных погрешностей является наиболее объективным, так как целиком базируется на теории случайных ошибок. Однако использование его при большом числе частных погрешностей без наложения ряда априорных заключений затруднительно.

К сожалению, пока нет сравнительных данных по расчету погрешностей приборов этими методами.

2.2.7. Нормирование погрешностей измерительных устройств

Погрешность измерительных устройств определяется на основании поверок и до настоящего времени сопровождается указанием предельного поля допуска $\pm\Delta_{\max}$, определенного для нормальных условий работы, который не должен быть превзойден в диапазоне изменения измеряемой величины в течение гарантированного времени. Такой подход к оценке погрешностей не может считаться правильным, если учесть, что погрешность измерительных устройств, как уже говорилось, определяется в основном случайными факторами и, следовательно, является случайной величиной. Это значит, что погрешность может принимать различные значения, большие и меньшие $\pm\Delta X_{\max}$, полученной на основании одной-двух реализаций характеристики прибора. Исчерпывающей характеристикой точности, как уже неоднократно указывалось, является закон распределения погрешностей прибора, преобразователя. Однако, включение в паспортные данные таблиц или графиков, отражающих закон распределения погрешностей данного измерительного средства,

громоздко и мало пригодно для практических целей, так как погрешности должны быть выражены в цифровой форме. Поэтому стандартом предлагается для нормирования погрешностей использовать: предельные значения систематической составляющей погрешности θ ; ее математическое ожидание m_θ и среднее квадратическое отклонение σ_θ ; предельное значение среднего квадратического отклонения σ_δ случайной составляющей погрешности и т. д. с указанием при необходимости функции распределения. Если функция распределения близка к нормальной, то для определения доверительного интервала предлагается пользоваться кривыми, изображенными на рис. 2.44.

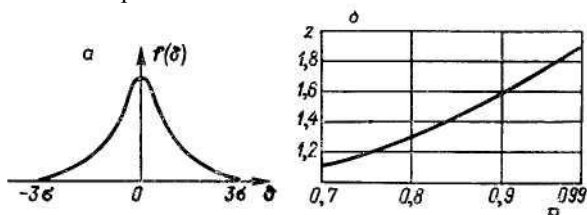


Рис. 2.44. Усеченный нормальный закон распределения. Зависимость z от P .

При этом интервал, в котором с заданной вероятностью P находится основная погрешность $\Delta_{тип}$ любого средства измерения данного типа при нормированных для нормальных условий значениях характеристик m_θ , σ_θ , σ_δ и допустимой вариации показаний b , определяется неравенством

$$m_\theta - z \sqrt{\sigma_\theta^2 + \frac{b^2}{12} + \sigma_\delta^2} < \Delta_{тип} < m_\theta + z \sqrt{\sigma_\theta^2 + \frac{b^2}{12} + \sigma_\delta^2}.$$

Стандарт распространяется на нормируемые метрологические характеристики, т. е. не только на погрешности в нормальных условиях, но и на ряд других важных характеристик (входной импеданс, функции влияния, динамические характеристики и т. д.). Выбор и установление нормируемых характеристик для измерительных средств разного типа, выбор формы нормирования — это задачи стандартов на различные типы измерительных средств.

Другая сторона вопроса нормирования погрешности касается формы нормирования: одночленной или двучленной формулой.

Стандарт допускает применение обеих формул, но не дает указаний, какая из них является предпочтительной в том или ином случае.

Вероятно, выразить погрешность одночленной формулой

(и соответственно нормировать) имеет смысл, когда аддитивная погрешность прибора (преобразователя) является определяющей и его мультипликативной ошибкой можно пренебречь. В противном случае необходимо нормировать двучленной формулой. В первом случае абсолютная погрешность

$$\Delta X = \pm a,$$

во втором

$$\Delta X = \pm (a + bX).$$

Постоянная a характеризует абсолютную аддитивную погрешность, b — относительную мультипликативную погрешность. Постоянные a , b подлежат нормированию, как бы они ни определялись: как предельные или как среднеквадратические.

Класс точности прибора определяется по относительной приведенной погрешности

$$\gamma_{\text{пр}} = \frac{\Delta X}{X_{\text{н}}} \cdot 100\%,$$

где $X_{\text{н}}$ — верхний предел измерения (преобразования) измерительного устройства. Для одночленной формулы

$$\gamma_{\text{пр}} = \pm \frac{a}{X_{\text{н}}} \cdot 100\%,$$

для двучленной (при $X = X_{\text{н}}$)

$$\gamma_{\text{пр}} = \pm \left(\frac{a}{X_{\text{н}}} \cdot 100\% + b \cdot 100\% \right).$$

Таким образом, и при нормировании двучленной формулой приведенную погрешность измерительного устройства можно выразить одним числом, которое определяет его класс. Однако это неудобно, так как аддитивная составляющая погрешности, проявляющаяся в смещении нуля, конечности квантов (резистивная разрешающая способность потенциометров, погрешность квантования в цифровых приборах и т. д.), не зависит от значения измеряемой величины, но определяет нижний предел измерения (преобразования) устройства. Поэтому целесообразнее класс точности прибора определять по величине b , т. е. по приведенной мультипликативной погрешности, а величину аддитивной погрешности указывать в абсолютных единицах (знаках делениях, в значениях измеряемой величины), как это и делают, например, для цифровых приборов. Тогда формулы нормирования принимают вид

$\gamma_s = \pm b \cdot 100\%$; $\Delta X_a = \pm a$, где γ_s — приведенная мультипликативная погрешность прибора.

При очень больших пределах измерения, когда верхний предел стремится к бесконечности, нормирование следует производить с помощью трехчленной формулы принимающей для абсолютной погрешности вид

$$\Delta X = \Delta X_a + \gamma_a X + X^2/X_m.$$

Так нормируются, например, погрешности омметров.

2.3. Погрешности измерительных устройств при изменении значения измеряемой величины во времени

2.3.1. Понятие динамической погрешности

Измеряемый сигнал, как отмечалось выше, не остается неизменным во времени, кроме тех случаев, когда он специально поддерживается постоянным. В табл. 2.1 приведены некоторые формы детерминированных сигналов, воздействий, с которыми часто приходится иметь дело при измерениях. Очевидно, что форма сигнала определяется процессом, параметр которого измеряется. Так, при наборе высоты летательным аппаратом (процесс) высота как параметр часто изменяется по линейному закону; при испытаниях на перегрузку различного рода устройств измеряемый сигнал обычно имеет вид импульса силы, при испытаниях на вибрацию измеряемый сигнал имеет вид синусоидальной функции и т. д. В общем случае сигнал может изменяться по законам, отличным от типовых, представленных в табл. 2.1, он может содержать низкочастотную и высокочастотную (детерминированные) составляющие, детерминированную и случайную составляющие или являться случайной функцией времени.

ИУ должны осуществлять преобразование входного воздействия в выходной сигнал, передаваемый в системы управления или регистрируемый специальными записывающими устройствами для последующего анализа без искажений, либо, если это невозможно, с искажениями, укладывающимися в определенные нормы. Выше было показано, что определение неизменного во времени сигнала производится ИУ с ошибками (только преобразователи, осуществляющие счет импульсов, оборотов и т. п., делают это без ошибки), которые часто для краткости называют *статическими*. Если же сигнал изменяется во времени, а ИУ обладает некоторой инерционностью, возникает еще одна ошибка — *динамическая*, причем часто большая, чем статическая погрешность.

Динамическая погрешность — это ошибка, которую правильнее было бы считать методической ошибкой измерения, так как величина

ее зависит не столько от ИУ, сколько от соотношений между параметрами сигнала и параметрами ИУ, т. е. определяется несоответствием ИУ поставленной задаче измерения сигнала $x(t)$. Это уже просчет экспериментатора, выбирающего аппаратуру, вольный или невольный. Как бы там ни было, ясно одно: для определения динамической погрешности измерения необходимо знать характеристики ИУ (преобразователя) и сигнала.

Как правило, динамическая погрешность интересует исследователей не сама по себе, так как она не определяет интервал, в котором с некоторой вероятностью лежит значение измеряемого сигнала (сигнал непрерывно меняется). Динамическую ошибку определяют обычно:

- для восстановления формы входного воздействия по выходному сигналу;
- для расчета корректирующих цепей к измерительным устройствам с целью улучшения их динамических характеристик;
- при исследовании влияния параметров ИУ (системы управления) на форму выходного сигнала при заданном входном воздействии.

Динамическая ошибка является функцией времени $\Delta_d(t)$, причем вид этой зависимости в общем случае может непрерывно меняться не только за счет изменения характера сигнала, что закономерно, но и за счет системы, если она относится к классу нестационарных. ИУ проектируются как стационарные системы, т. е. как системы, реакция которых на входной сигнал не зависит от времени его приложения и определяется только параметрами сигнала и ИУ. Тогда при постоянстве параметров ИУ вид $\Delta_d(t)$ будет определяться только $x(t)$. Если положить коэффициент преобразования ИУ равным единице, а статистические погрешности его равными нулю, то

$$\Delta_d(t) = y(t) - x(t) \quad (1.189)$$

На рис. 2.45 показано образование динамической ошибки преобразования линейного и синусоидального измеряемого сигнала линейным ИУ.

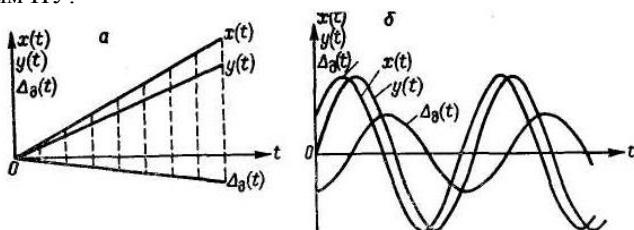


Рис. 2.45. К образованию динамической ошибки измерения при линейном (а) и синусоидальном (б) сигналах.

Оценку качества воспроизведения $x(t)$ измерительным устройством непосредственно по $\Delta_d(t)$, очевидно, производить неудобно. Поэтому для оценки качества воспроизведения должен быть принят некоторый критерий, позволяющий для целей данной задачи измерения дать наиболее объективную оценку точности. Разработка таких критериев и их выбор также составляют одну из задач, которые приходится решать при анализе динамики.

Принципиально, в чисто теоретическом плане, динамическая ошибка $\Delta_d(t)$ может быть найдена аналитически по (2.189), если одним из возможных методов математического анализа определено $y(t)$. Это может быть сделано путем интегрирования дифференциального уравнения, применением операционного метода, преобразования Фурье или метода интеграла Дюамеля. Практически это невозможно, так как входной сигнал аналитически не задан; он измеряется потому, что неизвестен. Однако во всех случаях измерений обычно имеется предварительно накопленная информация относительно характера сигналов, подлежащих измерению. Если такой информации нет, то иногда ее удается получить непосредственно перед измерением путем изучения явления, параметры которого измеряются. Чаще всего удается получить информацию о частотном спектре сигналов и о предельных значениях амплитуд, о скорости изменения отдельных составляющих сигналов, о предполагаемой длительности импульсов и т. д. Если подобная информация о сигнале получена, то можно из класса стандартизованных сигналов, приведенных в табл. 2.1, подобрать такой, который по некоторым показателям мог бы моделировать целое множество измерительных сигналов. Так, если известно, что сигнал периодический, то в качестве стандартизованного сигнала целесообразно выбрать синусоидальный, ибо любое периодическое воздействие может быть представлено рядом Фурье; если сигнал имеет вид импульса, то в качестве стандартизованного сигнала можно выбрать δ -функцию или единичное ступенчатое воздействие и т. д. Ниже этот вопрос будет рассмотрен достаточно подробно. Здесь же следует отметить, что если есть возможность выбора стандартизованного воздействия, то предпочтение отдается тому, которое проще может быть реализовано при необходимости экспериментальных исследований, проверок.

Таким образом, динамическую погрешность при теоретических и экспериментальных исследованиях можно оценить как

$$y_c(t) - x_c(t) \quad (2.190)$$

где $x_c(t)$ — стандартизованный сигнал, моделирующий измеряемый по некоторым показателям; $y_c(t)$ — реакция ИУ на $x_c(t)$.

По полученной $\Delta_{л.с}(t)$ и выбранному критерию оценки можно оценить качество воспроизведения любого измеряемого сигнала данного класса или восстановить $x(t)$ по виду выходного воздействия $y(t)$.

При проектировании ИУ задача ставится в несколько другом плане: найти оптимальные параметры ИУ, при которых выбранный критерий качества воспроизведения измеряемых сигналов данного класса был бы минимальным (задача синтеза).

2.3.2. Представление измеряемых сигналов стандартизованными воздействиями

Выбор стандартизованного сигнала $x_c(t)$ для изучения качества воспроизведения его заданной измерительной системой определяется, как уже указывалось, классом измеряемых сигналов, т. е. математической целесообразностью. Но не только класс сигналов определяет выбор $x_c(t)$, так как один и тот же сигнал может быть представлен различными стандартизованными воздействиями. Выбор $x_c(t)$ зависит также от формы представления динамических характеристик ИУ, преобразователя, от класса преобразователей. Так, преобразователи, представляющие собой колебательную систему, удобно описывать частотными характеристиками; преобразователи типа апериодического звена удобнее описывать временными характеристиками, и т. д.

Рассмотрим некоторые формы представления измеряемых сигналов.

1. $x(t)$ — периодическая функция произвольного вида, но не синусоидальная (рис. 2.46).

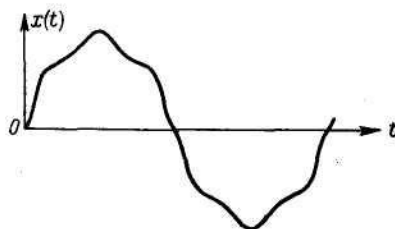


Рис. 2.46. Произвольный периодический процесс.

Если эта функция может быть разложена в сходящийся ряд Фурье, то

$$x(t) = B_0 + \sum_{k=1}^{\infty} A_k \sin k\omega_1 t + \sum_{k=1}^{\infty} B_k \cos k\omega_1 t, \quad (2.191)$$

где ω_1 — круговая частота первой гармоники; k — номер гармоники.

Коэффициенты разложения A_k, B_k имеют, как известно из курсов высшей математики, вид

$$A_k = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \sin k\omega_1 t dt; \quad (2.192)$$

$$B_k = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \cos k\omega_1 t dt; \quad (2.193)$$

$$B_0 = \frac{1}{T} \int_{-T/2}^{T/2} x(t) dt. \quad (2.194)$$

2. $x(t)$ — непериодическая функция произвольной формы.

Если $x(t)$ удовлетворяет условиям Дирихле и абсолютно интегрируема, то ее можно представить интегралом Фурье. Так как непериодическая функция $x(t)$ может рассматриваться как периодическая с периодом $T \rightarrow \infty$, то представление $x(t)$ интегралом Фурье может быть получено из (2.191), если в нем выразить синус и косинус через показательные функции по формулам Эйлера. Так как

$$\sin k\omega_1 t = \frac{\exp(jk\omega_1 t) - \exp(-jk\omega_1 t)}{2j}, \quad (2.195)$$

$$\cos k\omega_1 t = \frac{\exp(jk\omega_1 t) + \exp(-jk\omega_1 t)}{2}, \quad (2.196)$$

$$\begin{aligned} \text{то } x(t) &= \frac{1}{2j} \sum_{k=-\infty}^{\infty} (A_k + jB_k) \exp(jk\omega_1 t) - \\ &- \frac{1}{2j} \sum_{k=-\infty}^{\infty} (A_k - jB_k) \exp(-jk\omega_1 t) = \sum_{k=-\infty}^{\infty} C_k \exp(jk\omega_1 t), \end{aligned} \quad (2.197)$$

где C_k — амплитуда k -й гармоники. В зависимости от k C_k определяются формулами:

$$C_k = \frac{B_k - jA_k}{2}; \quad C_{-k} = \frac{B_k + jA_k}{2}; \quad C_{k=0} = B_0.$$

Для выражения C_k в общей форме подставим значения A_k, B_k, B_0 из (2.192), (2.193) и (2.194). Тогда амплитуда k -й гармоники принимает вид

$$C_k = \frac{1}{T} \int_{-T/2}^{T/2} x(t) [\cos k\omega_1 t - j \sin k\omega_1 t] dt =$$

$$= \frac{1}{T} \int_{-T/2}^{T/2} x(t) \exp(-jk\omega_1 t) dt. \quad (2.198)$$

умножив обе части равенства (2.198) на $T=2\pi/\omega_1$, получим

$$C_k \frac{2\pi}{\omega_1} = \int_{-T/2}^{T/2} x(t) \exp(-jk\omega_1 t) dt. \quad (2.199)$$

Обозначив

$$D_k = C_k T = C_k \frac{2\pi}{\omega_1} = \int_{-T/2}^{T/2} x(t) \exp(-jk\omega_1 t) dt, \quad (2.200)$$

найдем

$$C_k = \frac{1}{2\pi} D_k \omega_1. \quad (2.201)$$

Если период T конечен, то он — интервал между отдельными амплитудами, выраженными через $D_k=C_k T$ и образующими линейчатый спектр функции $x(t)$. $D_k = C_k T$ есть k -я спектральная линия спектра $x(t)$, который может для наглядности быть изображен так, как это показано на рис. 2.47, а. При увеличении $T=2\pi/\omega_1$ ω_1 уменьшается, но длина спектральных линий D_k остается прежней для тех же значений $\omega_1, 2\omega_1$ и т. д., как это следует из выражений (2.198) и (2.199).

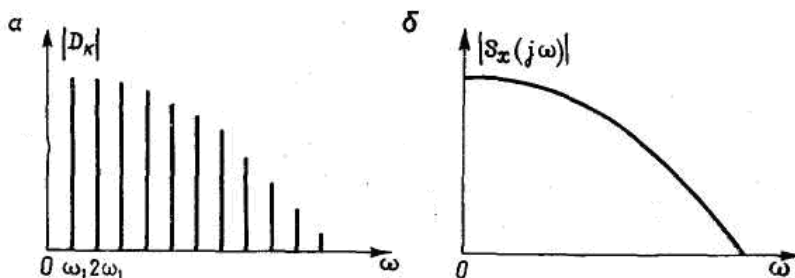


Рис. 2.47. К определению понятия спектральной плотности.

Пояснения в тексте.

Это значит, что уплотнение спектра происходит за счет появления новых линий. При $T \rightarrow \infty$ ($\omega_1 \rightarrow d\omega$) конечные точки спектральных линий сольются в непрерывную линию (рис. 2.47, б), представляющую

собой непрерывную функцию частоты ω и называемую комплексным частотным спектром или спектральной плотностью сигнала $x(t)$. Обозначают ее обычно через $S_x(j\omega)$. Из (2.200) при замене дискретных значений частоты текущей частотой ω и пределом интегрирования на $+\infty$ и $-\infty$ получим

$$S_x(j\omega) = \int_{-\infty}^{\infty} x(t) \exp(-j\omega t) dt. \quad (2.202)$$

Значения C_k в формуле (2.197) представления сигнала $x(t)$ при $T \rightarrow \infty$ будут определяться как

$$C_k = \frac{1}{\pi} S_x(j\omega) d\omega. \quad (2.203)$$

Подставляя (2.203) в (2.197) и заменяя знак суммы интегралом, найдем

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(j\omega) \exp(j\omega t) d\omega. \quad (2.204)$$

Таким образом, произвольная непериодическая функция может быть представлена бесконечной суммой бесконечно малых синусоидальных колебаний, содержащей все возможные частоты от 0 до ∞ , т. е. спектр непериодической функции сплошной в отличие от линейчатого спектра периодической функции. Модуль спектральной плотности $|S(j\omega)| = S(\omega)$ называется функцией распределения амплитуд спектра, или просто спектром, а $\psi(\omega)$ — функцией распределения начальных фазовых углов спектра, или фазовым спектром. Выражения (2.204) и (2.202) называются парой преобразований Фурье. Интеграл Фурье широко используется в практике измерений и для определения спектра измеряемых импульсов с целью оценки точности их воспроизведения. Так, спектральная плотность δ -функции

$$S_x(j\omega) = \int_{-\infty}^{\infty} \delta(t - \tau) \times \\ \times \exp(-j\omega t) dt = \exp(-j\omega \tau) \quad (2.205)$$

(модуль $S_x(\omega) = 1$).

Так как для $\delta(\tau) \tau \rightarrow 0$, то $S_x(j\omega) = 1$. Это означает, что δ -функция имеет сплошной спектр частот во всем возможном их диапазоне с неизменной спектральной плотностью.

Можно показать, что спектральная плотность любого короткого импульса равна постоянной величине, определяемой (приближенно) площадью импульса; спектр является сплошным и имеет границы частот, для которых справедливо соотношение

$$\tau \ll T, \quad (2.206)$$

где T — период, соответствующий граничной частоте.

Спектральная плотность прямоугольного импульса высотой h и длительностью τ (рис. 2.48)

$$S_x(j\omega) = h\tau \frac{\sin \omega\tau/2}{\omega\tau/2}. \quad (2.207)$$

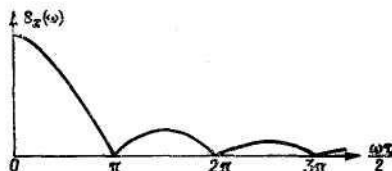


Рис. 2.48. Модуль спектральной плотности прямоугольного импульса.

Если неравенство (2.206) выполняется, то

$$\sin \frac{\omega\tau}{2} \approx \frac{\omega\tau}{2} \text{ и } S_x(j\omega) \approx h\tau. \quad (2.208)$$

Модули спектральных плотностей некоторых сигналов приведены в табл. 2.1.

Длительность реальных импульсов имеет конечные размеры, спектр их частот хотя и безграничен, но длина спектральных линий убывает достаточно быстро. Поэтому при анализе $x(t)$ нет смысла исходить из безграничности спектра частот. Можно условиться считать за ширину спектра полосу частот Δf между нулем и тем значением частоты, когда спектр первый раз обращается в нуль. Для прямоугольного импульса (см. рис. 2.48) спектр первый раз обращается в нуль при $\frac{\omega\tau}{2} = \pi$, т. е. при

$$2\pi\Delta f \frac{\tau}{2} = \pi, \quad (2.209)$$

откуда

$$\Delta f = 1/\tau. \quad (2.210)$$

Однако не всегда вопрос о выборе полосы частот может быть решен так однозначно, так как под длительностью импульса часто понимают разные величины: длительность при $h = 0$, длительность на высоте $0,5h$ и т. д. А. А. Харкевич предлагает под длительностью импульса понимать интервал времени, на который приходится подавляющая часть энергии импульса.

3. $x(t)$ —произвольное непериодическое воздействие, изображенное на рис. 2.49.

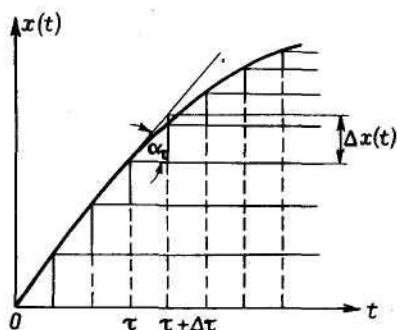


Рис. 2.49. Представление произвольной функции единичными.

Его можно представить в виде суммы ступенчатых воздействий типа единичной функции $1(t)$ или $1(t - \tau)$. Единичная функция (см. табл. 2.1) при $t < 0$ равна нулю; при $t \geq 0$ она скачкообразно становится равной единице и при всех $t > 0$ сохраняет это значение, что можно записать в виде

$$1(t) = \begin{cases} 0 & \text{при } t < 0, \\ 1 & \text{при } t \geq 0. \end{cases} \quad (2.211)$$

Единичная функция $1(t - \tau)$ запаздывает по отношению $1(t)$ на τ (рис. 2.50) и может быть выражена как

$$1(t - \tau) = \begin{cases} 0 & \text{при } t < \tau, \\ 1 & \text{при } t \geq \tau. \end{cases} \quad (2.212)$$

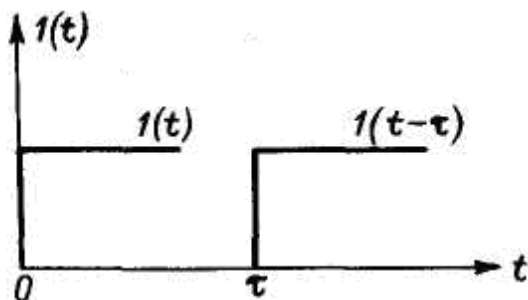


Рис. 2.50. Единичные функции.

Так как единичная функция не является абсолютно интегрируемой, то спектр ее непосредственно, как это делалось выше, определить нельзя. Для того чтобы найти аналитическое выражение единичной функции через спектр ее частот, можно воспользоваться искусственным приемом, а именно:

а) записать $1(t)$ в виде

$$1(t) = 1/2 \pm 1/2; \tag{2.213}$$

б) $\pm 1/2$ представить интегралом Дирихле :

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sin \omega t}{\omega} d\omega = \pm \frac{1}{2}; \tag{2.214}$$

в) подставить (2.214) в (2.213); тогда

$$1(t) = \frac{1}{2} + \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sin \omega t}{\omega} d\omega, \tag{2.215}$$

или

$$1(t) = \frac{1}{2} + \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\exp(j\omega t)}{j\omega} d\omega. \tag{2.216}$$

Из (2.216) следует, что для единичной функции

$$S_x(j\omega) = 1/j\omega; \quad S(\omega) = 1/\omega. \tag{2.217}$$

Вернемся теперь к представлению входного воздействия единичными функциями, т. е. ступеньками, запаздывающими одна относительно другой на τ , как это показано на рис. 2.49. Для качественного представления высота ступенек должна быть бесконечно малой. Выражая $\Delta x(t)$ в долях единичной функции, можно написать:

$$\frac{\Delta x(t)}{1(t-\tau)} = \text{tg } \alpha_{\tau} \Delta\tau,$$

или

$$\Delta x(t) = 1(t-\tau) \text{tg } \alpha_{\tau} \Delta\tau. \tag{2.218}$$

Переходя к пределам и заменяя tga значением производной $x(t)$ в точке $t=x$, получим:

$$dx(t) = 1(t-\tau) x'(\tau) d\tau; \tag{2.219}$$

$$x(t) = \int_0^t 1(t-\tau) x'(\tau) d\tau. \tag{2.220}$$

4. $x(t)$ —произвольная непериодическая функция, изображенная на рис. 2.51.

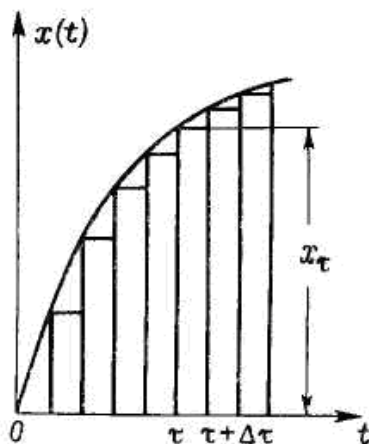


Рис. 2.51. Представление измеряемого процесса δ -функциями.

Ее можно представить последовательностью единичных импульсов в виде вертикальных столбиков, имеющих одинаковое основание $\Delta\tau$ и разную высоту $x(\tau)$. Выражая функцию импульса x_{Δ} в долях δ -функции, получим:

$$\frac{x_{\Delta}}{\delta(t - \tau)} = \frac{x(\tau)}{1/\Delta\tau} = x(\tau) \Delta\tau,$$

или

$$x_{\Delta}(t) = x(\tau) \delta(t - \tau) \Delta\tau. \quad (2.221)$$

Переходя к бесконечно малым и интегрируя в интервале времени, в котором функция $x(t)$ задана, получим

$$x(t) = \int_0^t x(\tau) \delta(t - \tau) dt. \quad (2.222)$$

Выражение (2.222), являющееся представлением произвольной функции $x(t)$ через δ -функцию, непосредственно вытекает из ее свойств: интеграл от произведения δ -функции на любую непрерывную функцию равен значению этой функции при том значении переменной интегрирования, при котором аргумент δ -функции $(t - \tau)$ обращается в нуль.

В заключение следует отметить, что δ -функция является производной от единичной функции

$$\delta(t) = \frac{d[1(t)]}{dt}, \quad (2.223)$$

или

$$1(t) = \int_{a \rightarrow 0}^t \delta(t) dt. \tag{2.224}$$

Это непосредственно следует из (2.211) с учетом обозначений, сделанных для этой функции в табл. 2.1.

2.3.3. Реакция измерительных устройств на синусоидально-изменяющийся входной сигнал

В общем случае работа ИУ описывается дифференциальным уравнением n -порядка

$$\begin{aligned} a_n \frac{d^n y(t)}{dt^n} + a_{n-1} \frac{d^{n-1} y(t)}{dt^{n-1}} + \dots + a_0 y(t) = \\ = b_m \frac{d^m x(t)}{dt^m} + b_{m-1} \frac{d^{m-1} x(t)}{dt^{m-1}} + \dots + b_0 x(t), \end{aligned}$$

где a_n, a_{n-1}, \dots, a_0 — коэффициенты уравнения, определяемые параметрами ИУ; b_m, b_{m-1}, \dots, b_0 — коэффициенты, определяемые параметрами сигнала.

На практике чаще встречается случай, когда в правой части уравнения все члены, кроме последнего, равны нулю. Тогда, положив $b_0 = B$, получим

$$a_n \frac{d^n y(t)}{dt^n} + a_{n-1} \frac{d^{n-1} y(t)}{dt^{n-1}} + \dots + a_0 y(t) = Bx(t), \tag{2.225}$$

или, в операторной форме:

$$y(t) [a_n p^n + a_{n-1} p^{n-1} + \dots + a_0] = Bx(t). \tag{2.226}$$

Обозначая

$$a_n p^n + a_{n-1} p^{n-1} + \dots + a_0 = Q(p), \tag{2.227}$$

найдем передаточную функцию

$$W(p) = B/Q(p) = \frac{B}{a_n p^n + a_{n-1} p^{n-1} + \dots + a_0}. \tag{2.228}$$

Число членов левой части (2.225), т. е. порядок уравнения, зависит от вида преобразователей, входящих в ИУ, от их динамических свойств. Каждый отдельный преобразователь принципиально также может быть описан уравнением k -го порядка ($k > 2$).

Однако для практических целей при описании отдельных преобразователей, как было показано в 2.2.1, можно ограничиться уравнением второго порядка, рассматривая преобразователи как

элементарные стационарные звенья, характеризующиеся передаточными функциями $W_i(p)$. В табл. 2.3.1 приведены передаточные функции некоторых элементарных стационарных звеньев. Зная структуру измерительной цепи и передаточные функции преобразователей, легко найти передаточную функцию ИУ. Для цепи прямого преобразования, изображенной на рис. 2.52, а, передаточная функция

$$W(p) = W_1(p) \cdot W_2(p) \cdot \dots \cdot W_i(p) \cdot \dots \cdot W_n(p), \quad (2.229)$$

для цепи уравнивающего преобразования (рис. 2.52, б)

$$W(p) = \frac{W_1(p)}{1 + W_1(p) \cdot W_2(p)}, \quad (2.230)$$

где $W_1(p)$ — передаточная функция цепи прямого преобразования; $W_2(p)$ — передаточная функция цепи обратного преобразования.

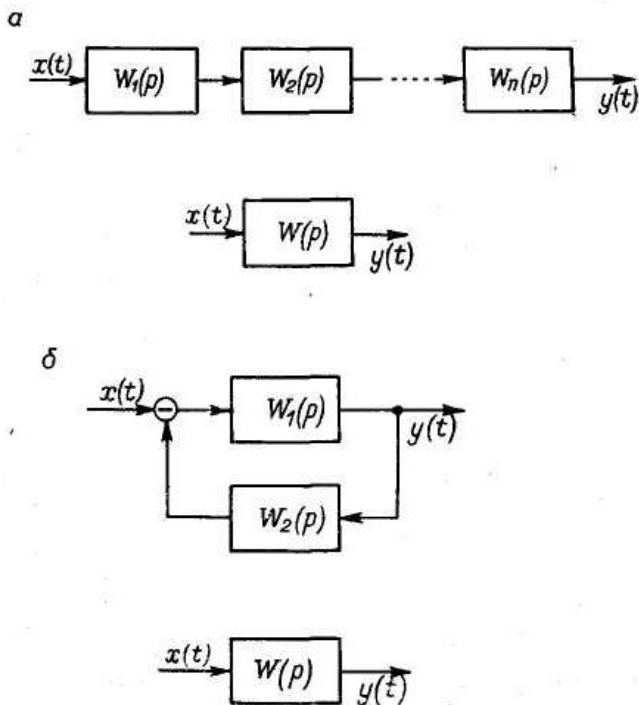


Рис. 2.52. Функциональные схемы измерительных устройств.
 а — прямое преобразование, б — уравнивающее.

В общем случае преобразователи могут быть объединены разными достаточно сложными структурными схемами.

Однако, используя методы преобразования цепей, хорошо разработанные в теории автоматического управления и регулирования, любую сложную схему можно преобразовать к виду, удобному для исследования. Так, схема на рис. 2.53, *а* может быть преобразована в схему рис. 2.53, *б*, где

$$k_1 = \frac{1}{2\beta T}; \quad T_1 = \frac{T}{2\beta}.$$

Определение выходного сигнала измерительного устройства путем решения (2.225) может быть сделано двумя способами: классическим методом интегрирования и операторным. Оба метода хорошо известны и широко применяются, особенно символический операторный метод для нахождения частного решения (2.225) при $x(t) = X_m \sin \omega t$ и определения частотных характеристик ИУ. Символический метод реализуется заменой оператора p в выражении для передаточной функции ИУ (2.228) на $j\omega$:

$$S(j\omega) = \frac{y(t)}{x(t)} = \frac{B}{a_n (j\omega)^n + a_{n-1} (j\omega)^{n-1} + \dots + a_0}, \quad (2.231)$$

где $S(j\omega) = W(j\omega)$ —комплексная чувствительность или комплексный коэффициент преобразования ИУ. Модуль этого выражения дает амплитудно-частотную характеристику, а угол сдвига — фазочастотную.

Рассмотрим измерительную систему, обобщенная схема которой изображена на рис. 2.21. Если измеряемое воздействие $x = X_m \sin \omega t$, то уравнение (2.155), (2.156) и (2.157) принимают вид

$$\frac{d^2 y(t)}{dt^2} + 2\beta \omega_0 \frac{dy(t)}{dt} + \omega_0^2 y(t) = \omega_0^2 S(0) X_m \sin \omega t; \quad (2.232)$$

$$y(t) [p^2 + 2\beta \omega_0 p + \omega_0^2] = \omega_0^2 S(0) X_m \sin \omega t; \quad (2.233)$$

$$W(p) = \frac{\omega_0^2 S(0)}{p^2 + 2\beta \omega_0 p + \omega_0^2}, \quad (2.234)$$

где $S(0)$ —чувствительность измерительного устройства к постоянному во времени измеряемому сигналу.

Таблица 2.3.1

Передаточные функции элементарных стационарных звеньев

Название звена	Уравнение звена	Передаточная функция	Весовая функция
Усилительное	$y(t) = kx(t)$	k	$k\delta(t)$
Запаздывающее	$y(t) = x(t - \tau)$	$\exp(-\tau p)$	$\delta(t - \tau)$
Интегрирующее	$y(t) = \int_{t_0}^t x(\tau) d\tau$	$\frac{1}{p}$	$1(t)$
Дифференцирующее	$y(t) = x'(t)$	p	$\delta'(t)$
Форсирующее первого порядка	$y(t) = Tx'(t) + x(t)$	$Tp + 1$	$T\delta'(t) + \delta(t)$
Апериодическое	$Ty'(t) + y(t) = Bx(t)$	$\frac{B}{Tp + 1}$	$\frac{k}{T} \exp\left(-\frac{t}{T}\right)$
Колебательное	$T^2y''(t) + 2\beta Ty'(t) + y(t) = Bx(t)$	$\frac{B}{T^2p^2 + 2T\beta p + 1}$	$\frac{k}{T\sqrt{1-\beta^2}} \exp\left(-\frac{\beta t}{T}\right) \sin\left(\frac{t}{T}\sqrt{1-\beta^2}\right)$
Резонансное	$T^2y''(t) + y(t) = Bx(t)$	$\frac{B}{T^2p^2 + 1}$	$\frac{k}{T} \sin \frac{t}{T}$

Примечание. Через T обозначена величина обратная круговой частоте собственных колебаний

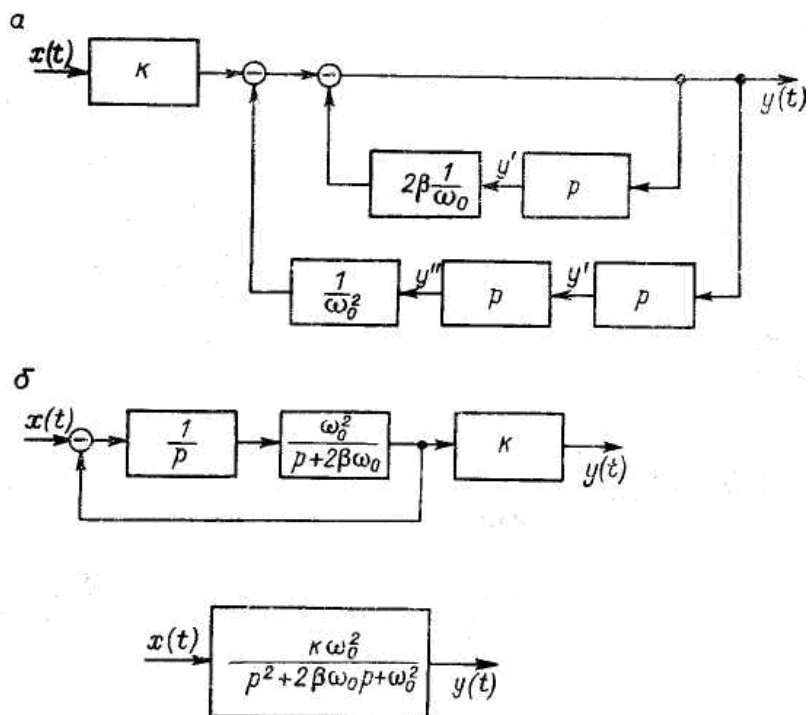


Рис. 2.53. Функциональные схемы колебательного звена.

Если в (2.234) p заменить на $j\omega$, получим комплексную чувствительность

$$S(j\omega) = \frac{\omega_0^2 S(0)}{(j\omega)^2 + 2\beta\omega_0 j\omega + \omega_0^2}, \quad (2.235)$$

или

$$S(j\eta) = \frac{S(0)}{(1 - \eta^2) + j2\eta\beta}, \quad (2.236)$$

где $\eta = \omega/\omega_0$ — отношение частоты измеряемого сигнала к частоте собственных колебаний.

Амплитудно-частотная характеристика

$$S(\eta) = \frac{1}{\sqrt{(1 - \eta^2)^2 + (2B\eta)^2}}, \quad (2.237)$$

фазочастотная

$$\psi(\eta) = \arctg \frac{2\beta\eta}{1-\eta^2}. \quad (2.238)$$

Частное решение уравнения (2.232):

$$y(t) = S(j\omega) \cdot x(t), \quad (2.239)$$

или

$$\begin{aligned} y(t) &= S(0) \cdot S(\eta) X_m \sin[\omega t - \psi(\eta)] = \\ &= \frac{S(0)}{\sqrt{(1-\eta^2)^2 + (2\beta\eta)^2}} X_m \sin\left(\omega t - \arctg \frac{2\beta\eta}{1-\eta^2}\right). \end{aligned} \quad (2.240)$$

Анализ (2.240) показывает, что чувствительность ИУ к синусоидальному воздействию не равна его чувствительности к постоянному возмущению и не является постоянной величиной, так как при $\beta = \text{const}$ зависит от отношения частот $\eta = \omega/\omega_0$. Только при $\omega_0 \gg \omega$ ($\eta \rightarrow 0$) чувствительность к синусоидальному возмущению

$$S(0) \cdot S(\eta) \rightarrow S(0). \quad (2.241)$$

То обстоятельство, что чувствительность $S(0) \cdot S(\eta)$ зависит от η , приводит к искажению формы кривой выходного сигнала, если входной сигнал является периодической функцией, содержащей синусоидальные гармоники разных частот. Искажение формы выходного сигнала происходит также за счет угла сдвига $\psi(\eta)$, который также будет разным для разных гармоник. В результате выходной сигнал как бы «размывается», как это показано на рис. 2.54 для периодического сигнала, содержащего первую и третью гармоники.

Динамическая ошибка воспроизведения

$$\Delta_n(t) = y(t) - x(t) \text{ при } S(0) = 1 \quad (2.242)$$

изменяется во времени и определяется как амплитудными, так и фазовыми искажениями; разделить $\Delta_n(t)$ на составляющие, соответствующие амплитудным и фазовым искажениям, не удастся. Можно лишь отдельно характеризовать степень искажений, проявляющихся в возникновении динамической ошибки. Для этой цели используются амплитудно-частотные и фазочастотные характеристики ИУ.

На рис. 2.55 приведено семейство амплитудно-частотных характеристик для различных значений степени успокоения β (значения η по оси абсцисс отложены в логарифмическом масштабе).

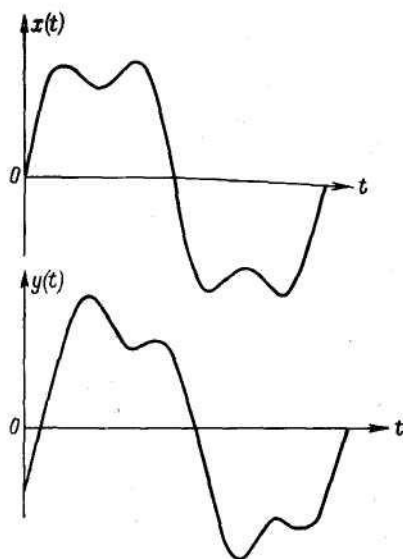


Рис. 2.54. Искажение сигнала измерительным устройством.

Выше уже отмечалось, что для правильного воспроизведения периодического воздействия, а также непериодического при представлении его интегралом Фурье (2.204) необходимо, чтобы чувствительность $S(0) \cdot S(\eta)$ и $\psi(\eta)$ оставались постоянными для спектра частот, которыми входное воздействие может быть представлено. Из рассмотрения представленных на рис. 2.55 кривых можно сделать вывод, что лучше всего такому требованию удовлетворяет амплитудно-частотная характеристика, соответствующая $\beta = 0,707$. Для нее при изменении в пределах от 0 до 0,3 отличие $S(t)$ от единицы составляет величину, меньшую 1%. Значение $\eta = 0$ соответствует либо $\omega = 0$, либо $\omega_0 \rightarrow \infty$. Первое тривиально, так как соответствует $x(t) = \text{const}$, второе нереально для системы, обладающей инерционностью. Соотношение $\eta = 0,3$ соответствует неравенству

$$\omega_0 > 3\omega, \quad (2.243)$$

$$\text{или } f_0 > 3f. \quad (2.244)$$

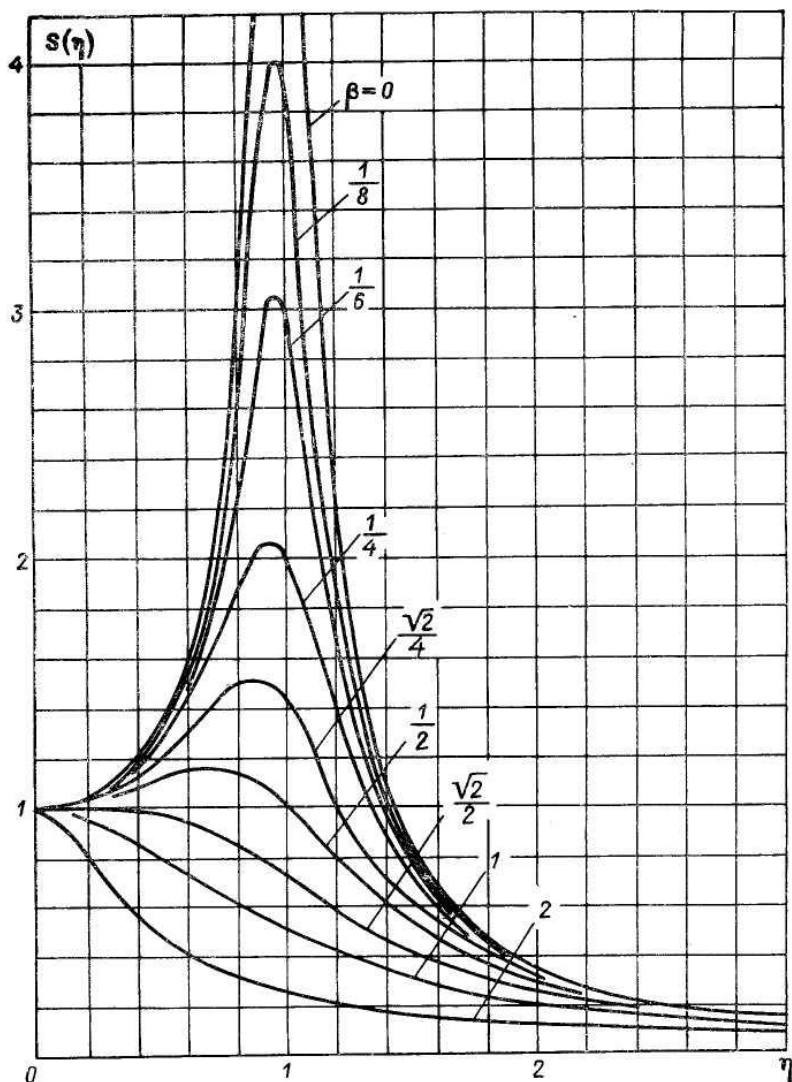


Рис. 2.55. Амплитудно-частотная характеристика колебательной системы

Чем лучше выполняется соотношение (2.244), тем меньше амплитудные искажения влияют на величину динамической ошибки. Частота собственных колебаний f_0 конкретного ИУ задана его конструкцией; частотный диапазон, т. е. диапазон частот измеряемого

сигнала, при котором амплитудные искажения не превысят заданной нормы, определяется как

$$f < kf_0. \tag{2.245}$$

Коэффициент k задается допустимой величиной частотной погрешности. Это должно быть сделано с учетом спектра входного сигнала, т. е. более правильно (2.245) записать в форме

$$f_i < kf_0, \tag{2.246}$$

принимая в качестве f_i частоту гармоники, которую желательно воспроизвести с заданной точностью. Сопоставление кривых, изображенных на рис. 2.55, показывает, что амплитудно-частотные характеристики очень критичны к степени успокоения β . Это следует учитывать при выборе (проектировании) измерительных средств.

Мы рассмотрели влияние амплитудных искажений на величину динамических искажений. Естественно, желательно получить аналогичные выводы о выборе η и β , исходя из фазовых искажений. На рис. 2.56 представлено семейство фазочастотных характеристик.

Фазовые искажения воспроизведения будут отсутствовать, если для всех гармоник процесса будет справедливо соотношение

$$\psi(\omega_k) = \omega_k t_{\text{пр}} = k\psi(\omega_1) = \text{const}, \tag{2.247}$$

где $t_{\text{пр}}$ — так называемое время пробега, прохождения всех составляющих спектра $x(t)$ через ИУ.

Это означает, что фазовый сдвиг k -й гармоники должен быть в k раз больше сдвига основной гармоники. Для доказательства (2.247) перепишем в виде

$$\frac{\psi(\omega_k)}{\omega_1} = kt_{\text{пр}} = \text{const} \tag{2.248}$$

и выразим гармоники в единицах ω_0 , как это делалось выше. Тогда для общего случая

$$\frac{\psi(\eta)}{\eta} = \text{const}, \tag{2.249}$$

или

$$\psi(\eta) = C\eta. \tag{2.250}$$

Это означает, что при выполнении условия (2.247) связь между $\psi(\eta)$ и η является линейной. Остается выяснить, при каких значениях η и β выполняется это условие. Из рассмотрения рис. 2.55 это выяснить не удается.

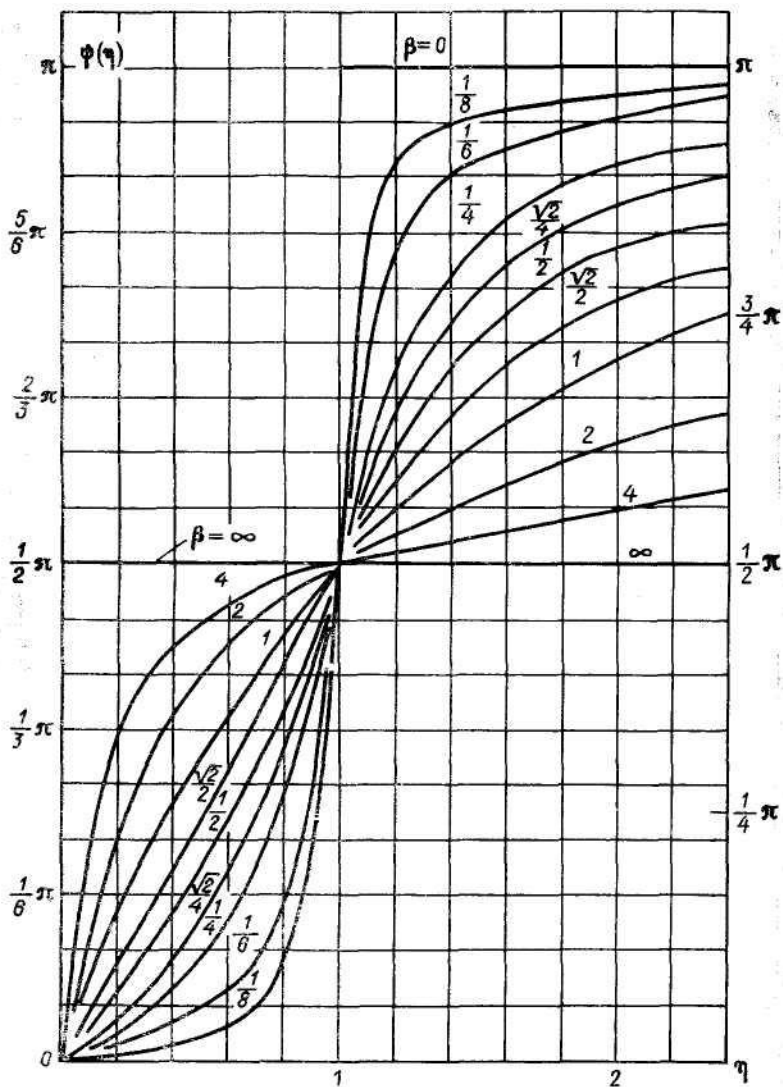


Рис. 2.56. Фазочастотные характеристики колебательной системы.

Ответ на поставленный вопрос дает семейство кривых

$$\frac{\psi(\eta)}{\eta} = f(\eta), \quad (2.251)$$

приведенное на рис. 2.57 для различных степеней успокоения.

Для $\beta = 0,75—0,85$ и $\eta \leq 1$ кривые фазочастотной погрешности почти совпадают с горизонтальной прямой, соответствующей $\pi/2$, т. е. для указанных диапазонов β и η фазовые искажения практически отсутствуют. Если сравнить условия минимума амплитудных и фазовых искажений, то можно видеть, что они несколько не совпадают. Для минимума амплитудных и фазовых искажений необходимо иметь $\eta \rightarrow 0$, т. е. $f_0 \gg \dot{f}$. С точки зрения амплитудных искажений это неравенство должно выполняться более строго, поэтому и следует принять (2.244) (2.246). Что касается степени успокоения β , то она получает разные численные значения при минимизации амплитудных и фазовых искажений: для минимизации амплитудных искажений β должно быть равно 0,707, фазовых — 0,75-0,85. Обычно принимают $\beta=0,707$, а удержание фазовых искажений в заданных нормах обеспечивают некоторым усилением неравенства (2.246).

Сделанные выводы относительно частотного диапазона ИУ, описываемых дифференциальным уравнением второго порядка, являются очень важными для практических целей. Дело в том, что ИУ, как ряд преобразователей, связанных структурной схемой, хотя принципиально и может иметь порядок уравнения $n > 2$, но часто может быть с достаточной степенью точности описано уравнением второго порядка, если произвести анализ его с точки зрения сопоставимости инерционных свойств преобразователей и скорости (частоты) изменения входного сигнала.

Мы рассмотрели возможность получения реакции ИУ на синусоидальное входное воздействие путем решения уравнения (2.232), описывающего работу ИУ. Аналогично решается и уравнение (2.225): также из комплексной чувствительности (2.231) можно определить амплитудно-частотную характеристику и угол сдвига. Однако сделать это значительно сложнее, чем для уравнения второго порядка. Для определения реакции на произвольное входное воздействие необходимо просуммировать реакции ИУ на все гармонические составляющие, которыми оно может быть представлено, независимо от того, как они получены: с помощью ряда или интеграла Фурье. При этом, так как спектр непериодической функции сплошной, то ограничиваются полосой частот $\Delta f = f_B - f_H$, о которой уже говорилось в предыдущем

пункте, т. е. рассматривают действие гармонических составляющих сигнала, лежащих в определенной полосе частот.

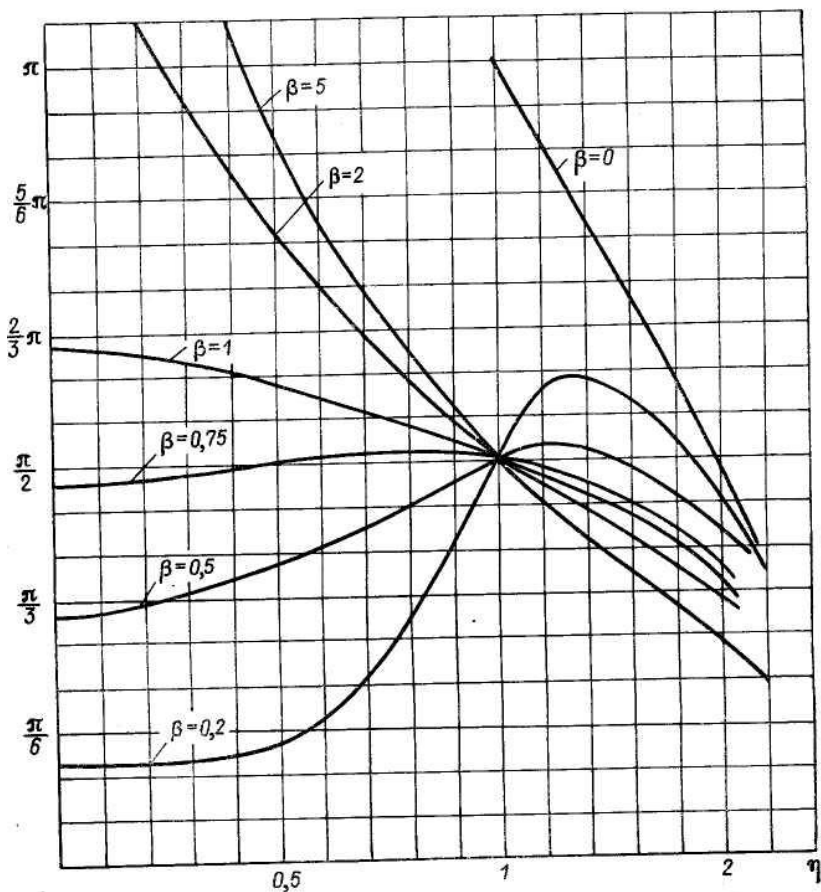


Рис. 2.57. Относительные фазочастотные характеристики.

Перейдем к анализу другого метода определения реакции ИУ на входное воздействие — метода интеграла Фурье, позволяющего непосредственно определить выходной сигнал при произвольном входном воздействии. Действительно, согласно (2.204) входной сигнал

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(j\omega) \exp(j\omega t) d\omega,$$

т. е. равен пределу суммы бесконечно малых элементарных составляющих

$$dx(t) = \frac{1}{2\pi} S_x(j\omega) \exp(j\omega t) d\omega, \quad (2.252)$$

имеющих амплитуду (по аналогии с $z = Z_m \exp(j\omega t) = Z_m \sin \omega t$)

$$dX_m = \frac{1}{2\pi} S_x(j\omega) d\omega, \quad (2.253)$$

где dX_m — амплитуда $dx(t)$; $S_x(j\omega)$ — спектральная плотность входного сигнала.

Бесконечно малая амплитуда выходного сигнала

$$dY_m = \frac{1}{2\pi} S(j\omega) \cdot S_x(j\omega) d\omega, \quad (2.254)$$

а выходной сигнал от действия $dx(t)$

$$dy(t) = S(j\omega) dx(t) = \frac{1}{2\pi} S(j\omega) \cdot S_x(j\omega) \exp(j\omega t) d\omega, \quad (2.255)$$

где $S(j\omega)$ — комплексная чувствительность ИУ, определяемая для общего случая по (2.231).

Для нахождения реакции ИУ на полное входное воздействие $x(t)$ выражение (2.255) необходимо проинтегрировать в пределах спектра частот. Так как спектр сплошной, то пределы интегрирования равны $+\infty$ и $-\infty$. На практике, как уже отмечалось, принимая сплошным спектр для произвольного входного воздействия, полосу частот ограничивают некоторым нижним f_n и верхним f_b значениями. Тогда в общем виде

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(j\omega) \cdot S_x(j\omega) \exp(j\omega t) d\omega, \quad (2.256)$$

а для практических целей

$$y(t) = \frac{1}{2\pi} \int_{\omega_n}^{\omega_b} S(j\omega) \cdot S_x(j\omega) \exp(j\omega t) d\omega. \quad (2.257)$$

Из последних уравнений следует, что спектральная плотность выходного сигнала

$$S_y(j\omega) = S(j\omega) \cdot S_x(j\omega) \quad (2.258)$$

отличается от спектральной плотности входного воздействия и величина этого отличия определяется частотными свойствами ИУ, представленными в (2.258) комплексной чувствительностью

Таким образом, зная спектральную плотность входного сигнала (2.202) и комплексную чувствительность (2.231), можно аналитически,

путем интегрирования (2.257), найти реакцию ИУ на произвольное входное воздействие $x(t)$.

В заключение следует напомнить, что полученные выражения для определения реакции ИУ на входное возмущение являются частными решениями уравнений (2.225) и (2.232). Общее решение этих уравнений, как известно, представляет собой линейную комбинацию показательных функций $C_1 \exp(z_1 t)$, $C_2 \exp(z_2 t)$, ..., $C_n \exp(z_n t)$, где z_1, z_2, \dots, z_n — корни характеристического уравнения

$$a_n z^n + a_{n-1} z^{n-1} + \dots + a_0 = 0, \quad (2.259)$$

а C_1, C_2, \dots, C_n — произвольные коэффициенты, определяемые из начальных условий.

При отсутствии кратных корней общее решение

$$y(t) = \frac{B}{a_n p^n + a_{n-1} p^{n-1} + \dots + a_0} \exp(pt) + C_1 \exp(z_1 t) + C_2 \exp(z_2 t) + \dots + C_n \exp(z_n t). \quad (2.260)$$

Следовательно, передаточная функция для любого момента времени

$$W^*(p) = \frac{y(t)}{x(t)} = \frac{B}{a_n p^n + a_{n-1} p^{n-1} + \dots + a_0} + C_1 \exp[(z_1 - p)t] + C_2 \exp[(z_2 - p)t] + \dots + C_n \exp[(z_n - p)t], \quad (2.261)$$

так как

$$x(t) = X_m \sin \omega t = X_m \exp(pt).$$

Если действительные части разностей $(z_1 - p)$, $(z_2 - p)$ и т. д. отрицательны (ИУ имеет должное успокоение), то все показательные функции в (2.261) при $t \rightarrow \infty$ стремятся к нулю, т. е. переходный процесс, вызванный изменением $x(t)$, затухает. Форма переходного процесса, определяемая составляющими вида $C_i \exp(z_i t)$ в (2.260), и скорость его затухания зависят от корней характеристического уравнения, т. е. от параметров ИУ.

В литературе приводится приближенное общее решение уравнения второго порядка при $x(t) = X_m \sin \omega t$. На рис. 2.58 представлена графическая интерпретация общего решения для различных соотношений частот.

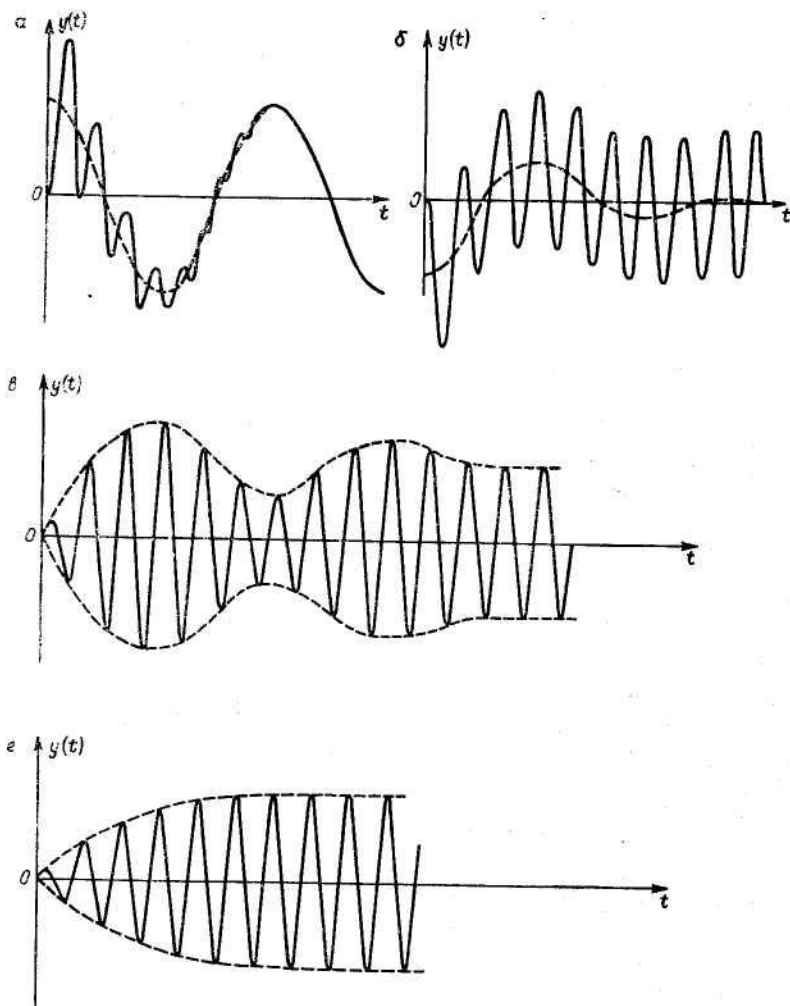


Рис. 2.58. Переходные процессы в колебательной системе при синусоидальном входном воздействии:

$$\left[\begin{array}{l} \alpha - \omega \ll \omega_0 \sqrt{1 - \beta^2}; \\ \beta - \omega \approx \omega_0 \sqrt{1 - \beta^2}; \end{array} \quad \begin{array}{l} \delta - \omega \gg \omega_0 \sqrt{1 - \beta^2}; \\ \varepsilon - \omega = \omega_0 \sqrt{1 - \beta^2}. \end{array} \right]$$

2.3.4. Реакция измерительных устройств на единичную функцию и δ -функцию

Найдем реакцию ИУ, описываемого дифференциальным уравнением второго порядка, на единичную функцию $1(t)$. Уравнение (2.232) при этом принимает вид

$$\frac{d^2 y(t)}{dt^2} + 2\beta\omega_0 \frac{dy(t)}{dt} + \omega_0^2 y(t) = \omega_0^2 S(0) \cdot 1(t). \quad (2.262)$$

Решение (2.262) проще всего получить классическим методом. Корни характеристического уравнения

$$z^2 + 2\beta\omega_0 z + \omega_0^2 = 0 \quad (2.263)$$

зависят от степени успокоения и могут быть выражены следующим образом:

$$\text{при } \beta < 1 \quad z_{1,2} = \omega_0(-\beta \pm j\sqrt{1-\beta^2}) = -\omega_0\beta \pm j\omega_0'; \quad (2.264)$$

$$\text{при } \beta > 1 \quad z_{1,2} = \omega_0(-\beta \pm \sqrt{\beta^2-1}) = -\omega_0\beta \pm \omega_0''; \quad (2.265)$$

$$\text{при } \beta = 1 \quad z_{1,2} = -\omega_0, \quad (2.266)$$

где

$$\omega_0' = \omega_0 \sqrt{1-\beta^2} \quad \text{и} \quad \omega_0'' = \omega_0 \sqrt{\beta^2-1}.$$

Для случая мнимых корней (2.264) решение надо искать в виде $y(t) = S(0) \cdot 1(t) + \exp(-\beta\omega_0 t) (C_1 \sin \omega_0' t + C_2 \cos \omega_0' t)$. (2.267)

Если при

$$t=0 \quad y(t)=0 \quad \text{и} \quad \frac{dy(t)}{dt}=0,$$

то постоянные интегрирования в (2.267) будут

$$C_2 = -S(0) \cdot 1(t); \quad C_1 = -S(0) \cdot 1(t) \frac{\beta}{\sqrt{1-\beta^2}}$$

и решение уравнения (2.262) принимает вид

$$y(t) = S(0) \cdot 1(t) \left[1 - \exp(-\beta\omega_0 t) \left(\frac{\beta}{\sqrt{1-\beta^2}} \sin \omega_0' t + \cos \omega_0' t \right) \right], \quad (2.268)$$

или (после несложных тригонометрических преобразований)

$$y(t) = S(0) \cdot 1(t) \left[1 - \frac{1}{\sqrt{1-\beta^2}} \exp(-\beta\omega_0 t) \sin(\omega_0'' t + \arctg \sqrt{\frac{1-\beta^2}{\beta^2}}) \right]. \quad (2.269)$$

На рис. 2.59 изображены входное единичное воздействие $1(t)$ и реакция $y(t)$ на него при $S(0) = 1$ (кривые 1 и 2 соответственно).

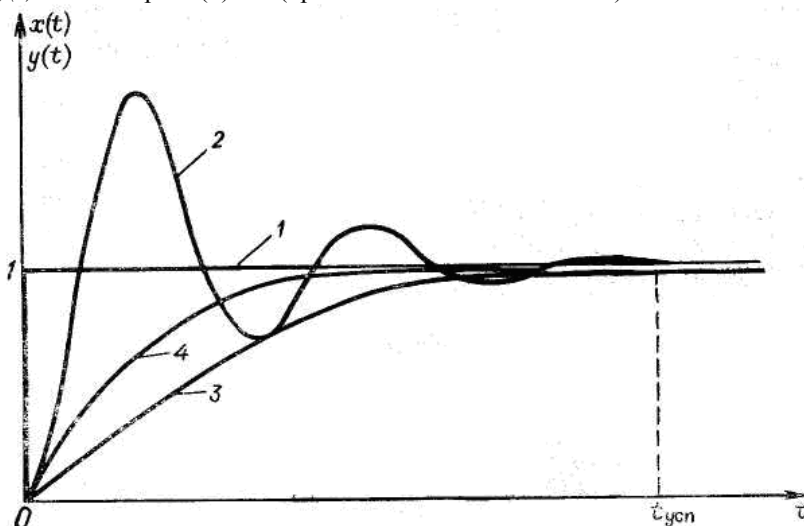


Рис. 2.59. Реакция колебательной системы на единичную функцию $1(t)$.

При $p > 1$ выходной сигнал выражается в аналогичной (2.269) форме, только вместо тригонометрических синуса и тангенса в формулу вводят синус и тангенс гиперболические, т. е.

$$y(t) = S(0) \cdot 1(t) \left[1 - \frac{1}{\sqrt{\beta^2-1}} \exp(-\beta\omega_0 t) \operatorname{sh}(\omega_0'' t + \operatorname{arcth} \sqrt{\frac{\beta^2-1}{\beta^2}}) \right]. \quad (2.270)$$

Наконец, при $\beta=1$, т. е. при наличии кратных корней, равных $-\omega_0$,

$$y(t) = S(0) \cdot 1(t) [1 - \exp(-\omega_0 t) (1 + \omega_0 t)]. \quad (2.271)$$

Кривые 3 и 4, отражающие реакцию ИУ при $\beta > 1$ и $\beta = 1$, приведены на том же рисунке.

Исследования показывают, что вторые слагаемые в (2.269), (2.270) и (2.271), характеризующие переходный процесс, становятся равными нулю при $t = \infty$. Это означает, что $y(t)$ достигнет уровня $x(t)$ при $S(0) = 1$ через бесконечно большое время, т. е. выходной сигнал $y(t)$ непрерывно меняется во времени вплоть до $t = \infty$, причем характер этих изменений зависит от степени успокоения. При $\beta < 1$ реакция ИУ на единичную функцию представляет собой колебательный процесс, амплитуда которого затухает по закону $\exp(-\omega_0 \beta t)$, где $\omega_0 \beta = \alpha$ — степень затухания. При $\beta > 1$ и $\beta = 1$ реакция ИУ на $1(t)$ есть аperiodический процесс, протекающий более или менее ускоренно в зависимости от $\omega_0 \beta$. При $\beta = 1$ $y(t)$ наилучшим образом (на первый взгляд) отражает входной сигнал. Этот режим работы ИУ называется критическим, граничным. При малейшем изменении по сравнению с единицей в любую сторону режим работы становится либо колебательным, либо аperiodическим. ИУ никогда не работают в критическом режиме — нереально выполнить при проектировании $\beta = 1$ из-за разброса параметров, определяющих β ; кроме того, β в реальных условиях работы изменяется от воздействия внешних факторов. Можно показать, что наиболее приемлемым, наиболее выгодным с различных точек зрения режимом работы является колебательный. Для этого, как уже говорилось выше, надо сформулировать критерий, что мы и сделаем ниже. Отношение

$$S(t) = h(t) = \frac{y(t)}{x(t)} \quad \text{при } x(t) = 1(t) \quad (2.272)$$

называется *переходной функцией или переходной чувствительностью*.

Переходная функция обычно обозначается через $h(t)$. Так как по отношению к ИУ $h(t)$ есть чувствительность, будем ее иногда обозначать $S(t)$.

Для измерительной системы второго порядка

$$S(t) = S(0) \left[1 - \exp(-\beta \omega_0 t) \sin \left(\omega_0' t - \arctg \sqrt{\frac{1 - \beta^2}{\beta^2}} \right) \right] \quad (2.273)$$

Переходная функция для системы n -го порядка находится аналогично тому, как это было сделано для системы второго порядка, и в общем виде может быть записана как

$$S(0) = S(0) [1 + C_1 \exp(z_1 t) + C_2 \exp(z_2 t) + \dots + C_n \exp(z_n t)] \quad (2.274)$$

где z_1, z_2, \dots, z_n — корни характеристического уравнения (не кратные); C_1, C_2, \dots, C_n — постоянные интегрирования, определяемые по начальным условиям.

Основная сложность заключается в определении корней характеристического уравнения. Если на входе действует $x(t)=1(t-\tau)$, т. е. момент начала измерения не совпадает с моментом подачи сигнала (рис. 2.60), то переходная характеристики, определяемая сигналом и ИУ, будет (2.273), так как в интервале от $t=0$ до $t=\tau$ $y(t)=0$.

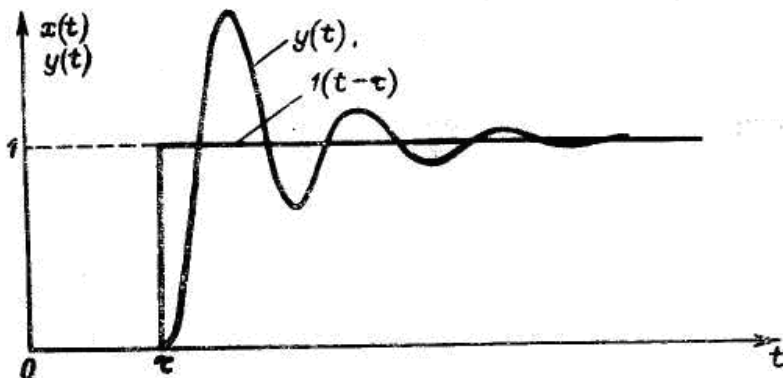


Рис. 2.60. Реакция колебательной системы на единичную функцию $1(t-\tau)$.

Однако, так как $S(t)$ имеет смысл только начиная с $t=\tau$, ее обозначают $S(t-\tau)$.

Определим реакцию ИУ на произвольное входное воздействие при представлении его серией единичных функций, имеющих неограниченную длительность, т. е. при представлении $x(t)$ интегралом

$$x(t) = \int_0^t x'(\tau) \cdot 1(t-\tau) d\tau. \quad (2.275)$$

Найдем вначале реакцию ИУ на бесконечно малую ступеньку

$$dx(t) = x'(\tau) \cdot 1(t-\tau) d\tau. \quad (2.276)$$

Для этого $dx(t)$ надо умножить на $S(t)$, определяемое уравнениями (2.273), (2.274). Тогда

$$dy(t) = S(t) \cdot dx(t) = x'(\tau) \cdot 1(t-\tau) \cdot S(t) d\tau, \quad (2.277)$$

или, так как

$$S(t) \cdot 1(t-\tau) = S(t-\tau) = h(t-\tau),$$

то

$$dy(t) = x'(\tau) \cdot S(t-\tau) d\tau. \quad (2.278)$$

Интегрируя последнее выражений, получим

$$y(t) = \int_0^t x'(\tau) \cdot S(t - \tau) d\tau. \quad (2.279)$$

Если в момент $t=0$ $x(t)$ имеет некоторое значение $x = x_1 \cdot 1(t) = \text{const}$ (рис. 2.61), то (2.275) следует переписать в виде

$$x(t) = x_1 \cdot 1(t) + \int_0^t x'(\tau) \cdot 1(t - \tau) d\tau. \quad (2.280)$$

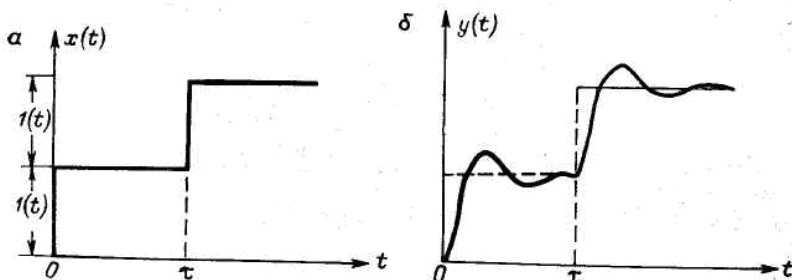


Рис. 2.61. Переходный процесс при ступенчатом изменении сигнала.
Пояснения в тексте.

Соответственно выходной сигнал

$$y(t) = x_1 \cdot 1(t) S(t) + \int_0^t x'(\tau) S(t - \tau) d\tau. \quad (2.281)$$

Выражения (2.279) и (2.281) носят названия интегралов Дюамеля и могут принимать также другие формы. Последнее является более общим, так как позволяет учесть влияние на выходной сигнал системы воздействия предыдущей ступеньки. Действительно, только в том случае, если τ на рис. 2.61, a равно бесконечности, т. е. второе изменение измеряемой величины наступило после того, как $y(t)$ к моменту $t = \tau$ приняло значение

$$y(t) = y_1 = x_1 \cdot 1(t) \cdot S(0), \quad (2.282)$$

можно считать, что

$$y(t) = y_1 + \int_0^t x_2'(\tau) \cdot S(t - \tau) d\tau, \quad (2.283)$$

как это показано на рис. 2.61, б.

При плавном изменении измеряемого сигнала скорость $x'(\tau)$ всегда меньше, чем при ступенчатом воздействии. Поэтому выходной сигнал будет лучше воспроизводить входное воздействие. Следовательно, переходная функция ИУ является некоторой предельной характеристикой, определяющей инерционные свойства ИУ. Поэтому

исследование динамических свойств ИУ с помощью единичной функции широко используется на практике, тем более, что и при экспериментальных исследованиях реализации $1(t)$ часто проще, чем реализация какого-либо другого детерминированного сигнала. Можно отметить еще одно достоинство исследований систем с помощью единичной функции. Дело в том, что производная от единичной функции есть δ -функция, т. е.

$$\delta(t) = \frac{d}{dt} [1(t)]. \quad (2.284)$$

Реакция системы на δ -функцию, называемая *импульсной или весовой функцией*, может быть определена как

$$g(t) = \frac{d}{dt} [h(t)] = \frac{d}{dt} [S(t)]. \quad (2.285)$$

Весовая функция, как реакция ИУ на δ -функцию, является исчерпывающей его характеристикой, так как δ -функция имеет сплошной спектр, простирающийся до бесконечно больших частот с неизменной спектральной плотностью (см. табл. 2.1)

$$S_{\delta}(j\omega) = \int_{-\infty}^{+\infty} \delta(t - \tau) \exp(-j\omega t) dt = \exp(-j\omega\tau), \quad (2.286)$$

модуль которой равен единице. При $\tau \rightarrow 0$ $S(j\omega) = 1$ и $\psi(\omega) = 0$.

Реакция ИУ на произвольное входное воздействие

$$x(t) = \int_{-\infty}^{+\infty} x(\tau) \delta(t - \tau) d\tau \quad (2.287)$$

может быть представлена как

$$y(t) = S \cdot x(t) = \int_{-\infty}^{+\infty} x(\tau) S_t \cdot \delta(t - \tau) d\tau, \quad (2.288)$$

где $S = \frac{y(t)}{x(t)}$ — оператор, или чувствительность ИУ. Индекс t у чувствительности S , означает, что ИУ реагирует на $x(t)$, выраженное через $\delta(t - \tau)$ в момент $\tau = t$. Следовательно, реакция ИУ на δ -функцию зависит от t и τ , т. е. от места расположения δ -импульса, которыми представляется $x(t)$ (см. рис. 2.49). Поэтому весовая функция

$$g(t, \tau) = g(t - \tau) = S_t(\delta - \tau). \quad (2.289)$$

С учетом введенного обозначения (2.289)

$$y(t) = \int_{-\infty}^{+\infty} x(\tau) \cdot g(t - \tau) d\tau. \quad (2.290)$$

Весовые функции некоторых элементарных стационарных звеньев приведены в табл. 2.3.1.

2.3.5. Связь между переходными, импульсными и частотными характеристиками измерительных устройств

В предыдущих параграфах было показано, что одно и то же *входное воздействие* может быть представлено различным образом: *в виде спектра, гармоник, единичными ступеньками или δ-функциями*. Выходной сигнал ИУ в (2.257), (2.281), (2.290) содержит множитель, отражающий свойства измеряемого сигнала. При представлении измеряемого сигнала *спектром* — это *спектральная плотность измеряемого сигнала*; при представлении $x(t)$ *единичной функцией* — это *производная от измеряемого сигнала* $x(t)$; при представлении $x(t)$ *через $\delta(t)$* — это *измеряемый сигнал*. Так как существует вполне определенная связь между $1(t)$, $\delta(t)$ и их спектрами, то обязательно должны существовать строгие соотношения между комплексной и переходной чувствительностями и весовой функцией, являющейся импульсной чувствительностью. Это формально, иначе реакция будет разной на одно и то же воздействие.

Связь между $S(j\omega)$, $S(t)=h(t)$, $S_u(t)=g(t)$ можно установить аналитически. Если в качестве входного воздействия $x(t)$ выбрать $\delta(t)$, то реакция ИУ согласно (2.256) будет

$$y(t) = g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(j\omega) \cdot S_{\delta}(j\omega) \exp(j\omega t) d\omega, \quad (2.291)$$

где $S_{\delta}(j\omega)$ — спектральная плотность δ -функции. Так как $S_{\delta}(j\omega)$ δ -функции равна единице, то

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} S(j\omega) \exp(j\omega t) d\omega. \quad (2.292)$$

Если сравнить полученное выражение с интегралом Фурье для произвольной функции

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} S_z(j\omega) \exp(j\omega t) d\omega, \quad (2.293)$$

где $S_z(j\omega)$ — спектральная плотность функции $f(z)$, то, очевидно, надо сделать вывод, что $S(j\omega)$ в (2.292) имеет смысл спектральной плотности выходной функции, т. е. весовой функции $g(t)$; наоборот, спектральная плотность весовой функции есть комплексная чувствительность

$$S(j\omega) = S_g(j\omega) = \int_{-\infty}^{\infty} g(t) \exp(-j\omega t) dt. \quad (2.294)$$

Таким образом, по известной весовой функции можно найти частотные характеристики ИУ. Наоборот, если известны частотные характеристики, можно определить весовую функцию ИУ.

Связь между весовой и переходной функциями определяется, как уже выше указывалось, соотношением

$$h(t) = \int_0^t g(\tau) d\tau. \quad (2.295)$$

Подставляя значение $g(\tau)$ из (2.292), получим

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{S(j\omega)}{j\omega} \exp(j\omega t) d\omega, \quad (2.296)$$

где $S_h(j\omega) = S(j\omega)/j\omega$ — спектральная плотность $h(t)$. Наоборот, спектральная плотность переходной функции

$$S_h(j\omega) = \int_{-\infty}^{+\infty} h(t) \exp(-j\omega t) dt. \quad (2.297)$$

При определении $h(t)$ и $g(t)$ через передаточную функцию необходимо в выражениях (2.296) и (2.294) $j\omega$ заменить через $p = \sigma + j\omega$. Тогда

$$h(t) = \frac{1}{2\pi j} \int_{\sigma - j\infty}^{\sigma + j\infty} \frac{W(p)}{p} \exp(pt) dp \quad (2.298)$$

$$\text{и } g(t) = \frac{1}{2\pi j} \int_{\sigma - j\infty}^{\sigma + j\infty} W(p) \exp(pt) dp. \quad (2.299)$$

Таким образом, для анализа динамических свойств ИУ достаточно знать какую-либо одну его функцию.

2.3.6. Реакция измерительного устройства на линейно-возрастающую во времени функцию

Линейно-возрастающая во времени функция относится к произвольным непериодическим воздействиям, и, следовательно, реакция ИУ на ее действие может быть определена одним из изложенных выше методов. Однако это можно сделать и классическим методом интегрирования уравнения или операционным. Попробуем выяснить, какой метод является более простым. Пусть

$$x(t) = vt \quad (2.300)$$

Изображение $x(t)$ по Лапласу

$$L\{x(t)\} = \int_0^{\infty} vt \exp(-pt) dt = v/p^2. \quad (2.301)$$

Если передаточная функция ИУ, например, имеет вид

$$W(p) = \frac{1}{Tp + 1}, \quad (2.302)$$

$$\begin{aligned} \text{то } y(t) &= L^{-1} \left\{ \frac{v}{p^2(Tp+1)} \right\} = \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{v}{p^2(Tp+1)} \exp(pt) dp. \end{aligned} \quad (2.303)$$

Решение легко найти при помощи вычетов. Подынтегральная функция (2.303) имеет два полюса: один полюс второго порядка ($p_1=0$) и один полюс первого порядка ($p_2=-1/T$). Вычет второго порядка

$$\begin{aligned} \text{Res}(p_1) &= \lim_{p \rightarrow p_1} \frac{d}{dp} [V(p)(p-p_1)^2] = \\ &= \lim_{p \rightarrow 0} \frac{d}{dp} \left[\frac{v \exp(pt)}{p^2(1+Tp)} p^2 \right] = vt - vT, \end{aligned} \quad (2.304)$$

где $V(p)$ —подынтегральная функция в (2.303). Вычет первого порядка

$$\begin{aligned} \text{Res}(p_2) &= \lim_{p \rightarrow p_2} [V(p)(p-p_2)] = \\ &= \lim_{p \rightarrow -1/T} \left[\frac{v \exp(pt)}{p^2(1+Tp)} (p+1/T) \right] = \\ &= \lim_{p \rightarrow -1/T} \left[\frac{v \exp(pt)}{p^2 T} \right] = vT \exp(-t/T). \end{aligned} \quad (2.305)$$

Тогда решение интеграла (2.303) будет

$$y(t) = v \{t - T[1 - \exp(-t/T)]\}. \quad (2.306)$$

На рис. 2.62 изображены входной и выходной сигналы для рассмотренного примера. Выходной сигнал запаздывает по отношению к входному с самого момента его приложения. Со временем величина отставания выходного сигнала становится практически постоянной, равной T , т. е. выходной сигнал с запаздыванием, но без заметных искажений следит за измеряемым сигналом. Следует заметить, что постоянная времени T , являющаяся характеристикой динамических свойств ИУ, не дает непосредственного представления о времени выхода $t_{\text{вых}}$ ИУ на квазистационарный режим, т. е. режим, для которого несовпадение $y(t)$ с пунктирной линией $y_1(t) = T + vt$ становится пренебрежимо малым. Для определения $t_{\text{вых}}$ следует задаться значением $\Delta y(t) = y(t) - y_1(t)$, при котором можно принять, что система вошла в практически стационарный режим. Пусть

$$\Delta y(t) = 0,01 y_m \cong 0,01 y_{1m}.$$

Тогда

$$0,01y_{1m} = vT \exp(-t_{\text{вых}}/T), \quad (3.307)$$

откуда

$$t_{\text{вых}} = T \ln \frac{vT}{0,01y_{1m}}. \quad (2.308)$$

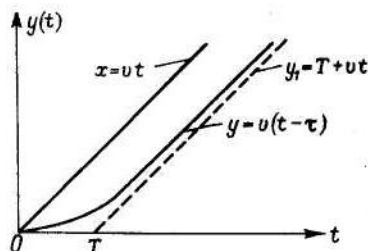


Рис. 2.62. Выходной сигнал аperiodического звена при линейно-возрастающем входном сигнале.

Время выхода $t_{\text{вых}}$ является очень важной характеристикой системы (критерием), особенно в случае, когда $x(t)$ есть кусочно-линейная функция, изображенная на рис. 2.63.

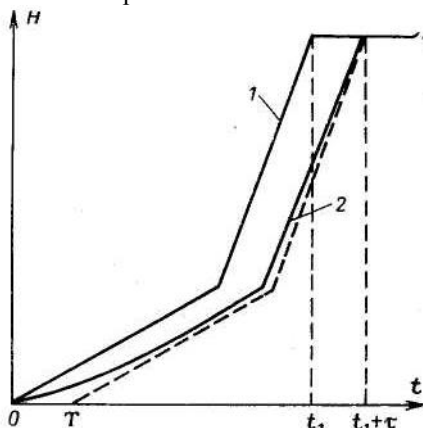


Рис. 2.63. Набор высоты.

1 — кривая действительного изменения высоты; 2 — показания высотомера (без коррекции).

В этом случае $t_{\text{вых}}$ обязательно должно быть меньше длительности участков кривой $x(t)$.

Если измеряемый сигнал $x(t)=vt$ представить δ -функцией, то

$$x(t) = \int_0^t v\tau\delta(t - \tau) d\tau, \quad (2.309)$$

а выходной сигнал (2.291)

$$y(t) = \int_0^t v\tau g(t - \tau) d\tau. \quad (2.310)$$

Весовая функция аperiodического звена, имеющего передаточную функцию (2.302),

$$g(t - \tau) = \frac{1}{T} \exp\left(-\frac{t - \tau}{T}\right). \quad (2.311)$$

Подставляя (2.311) в (2.310), получим

$$y(t) = \frac{v}{T} \int_0^t \tau \exp\left(-\frac{t - \tau}{T}\right) d\tau. \quad (2.312)$$

Это табличный интеграл, дающий при подстановке пределов

$$y(t) = v [t - T (1 - \exp(-t/T))],$$

совпадающее с (2.306). Последнее решение проще.

2.3.7. Оценка точности воспроизведения измерительными устройствами изменяющихся во времени сигналов

Динамическая погрешность измерения сигнала $x(t)$ является функцией времени и может быть представлена в виде

$$\Delta_d(t) = y(t) - S(0) \cdot x(t) = y(t) - x(t), \quad (2.313)$$

если $S(0)$ —чувствительность к $x(t)=\text{const}$ равна 1. Такое представление динамической ошибки, очевидно, ничего не дает, так как $x(t)$ — измеряемый неизвестный процесс.

Если задача определения динамической ошибки решается во временной области, то

$$\Delta_d(t) = \int_0^\infty x'(\tau) h(t - \tau) d\tau - h(0) x(t), \quad (2.314)$$

или

$$\Delta_d(t) = \int_0^\infty x(\tau) \cdot g(t - \tau) d\tau - g(0) x(t), \quad (2.315)$$

где $h(0)$ и $g(0)$ —переходная и весовая функции ИУ при $x(t) = \text{const}$.

При решении в частотной области

$$\Delta_d(t) = \frac{1}{\pi} \int_0^{\infty} S(j\omega) \cdot S_x(j\omega) \exp(j\omega t) d\omega - \\ - \frac{1}{\pi} \int_0^{\infty} S(0) \cdot S_x(j\omega) \exp(j\omega t) d\omega. \quad (2.316)$$

Таким образом, задача определения динамической ошибки может быть решена, если известна передаточная функция ИУ и спектральная плотность измеряемого сигнала. Так как между спектральными характеристиками, весовыми и переходными функциями существует вполне определенная связь, то расчет $\Delta_d(t)$ принципиально может быть выполнен по любой из формул (2.314), (2.315), (2.316). Измеряемый сигнал обычно задается спектральной плотностью. Поэтому проще воспользоваться последним уравнением, а первые два использовать для оценки точности воспроизведения по реакциям на стандартизированные возмущения $l(t)$ и $\delta(t)$.

Уравнение (2.316) можно получить в другом виде, если слагаемые в (2.313) выразить через передаточные функции:

$$\Delta_d(t) = [W(p) - W(0)] x(t). \quad (2.317)$$

Так как передаточная функция ИУ по ошибке

$$W_{\Delta}(p) = \frac{\Delta_d(p)}{x(p)}, \quad (2.318)$$

то формально $\Delta_d(t)$ может рассматриваться как выходной сигнал ИУ, имеющего передаточную функцию $W_{\Delta}(p)$, при воздействии на него $x(t)$. Обозначим $\Delta_d(t)$ через $y_{\Delta}(t)$. Тогда

$$y_{\Delta}(t) = \Delta_d(t) = W_{\Delta}(p) \cdot x(t) = S_{\Delta}(j\omega) \cdot x(t), \quad (2.319)$$

где $S_{\Delta}(j\omega)$ —комплексная чувствительность ИУ по ошибке.

Если измеряемый сигнал задан своим спектром

$$S_x(j\omega) = \int_0^{\infty} x(t) \exp(-j\omega t) dt, \quad (2.320)$$

то для нахождения $y_{\Delta}(t)$ можно воспользоваться методом нахождения интеграла Фурье, который был подробно описан в 2.3.3.

В соответствии с (2.256) и (2.258)

$$y_{\Delta}(t) = \frac{1}{\pi} \int_0^{\infty} S_{\Delta}(j\omega) \cdot S_x(j\omega) \exp(j\omega t) d\omega \quad (2.321)$$

и спектральная плотность ошибки

$$S_{\Delta\Delta}(j\omega) = S_{\Delta}(j\omega) \cdot S_x(j\omega), \quad (2.322)$$

или

$$S_{\Delta\Delta}(j\omega) = S_x(j\omega) \cdot [S(j\omega) - S(0)]. \quad (2.323)$$

Динамическая ошибка

$$\Delta_d(t) = y_{\Delta}(t) = \frac{1}{\pi} \int_0^{\infty} S_{\Delta\Delta}(j\omega) \exp(j\omega t) d\omega. \quad (2.324)$$

Для оценки точности воспроизведения можно воспользоваться средним квадратическим отклонением, равным

$$\sigma = \sqrt{\int_0^{\infty} [\Delta_d(t)]^2 dt}, \quad (2.325)$$

если $\Delta_d(t)$ абсолютно интегрируема.

Если оценка производится на конечном интервале времени τ , то

$$\sigma = \sqrt{\frac{1}{\tau} \int_0^{\tau} [\Delta_d(t)]^2 dt}. \quad (2.326)$$

Интегрирование квадрата функции по времени в пределах от 0 до ∞ можно заменить интегрированием квадрата модуля спектральной плотности этой функции по всем частотам, т. е. (2.325) выразить в виде

$$\sigma = \sqrt{\frac{1}{\pi} \int_0^{\infty} |S_{\Delta\Delta}(j\omega)|^2 d\omega}. \quad (2.327)$$

Обычно в качестве оценки точности называют величину

$$I = \sigma^2 = \int_0^{\infty} [\Delta_d(t)]^2 dt, \quad (2.328)$$

равную квадрату среднего квадратического отклонения. Такая оценка является наиболее объективной (по сути дела она характеризует качество передачи мощности сигнала через ИУ). При проектировании ИУ, выборе режимов их работы минимум I принимается в качестве критерия, т. е. параметры ИУ и режимы их работы выбираются из условия минимизации выбранного критерия.

В качестве критерия могут быть выбраны и другие оценки точности воспроизведения сигналов, например, интегральные оценки вида

$$I_1 = \int_{\delta}^{\infty} [\Delta_n(t)] dt \quad (2.329)$$

и

$$I_2 = \int_0^{\infty} |\Delta_n(t)| dt. \quad (2.330)$$

Первая пригодна только для монотонных процессов, вторая более совершенна, но вычисление модуля $\Delta_n(t)$, характеризующейся суммой членов, число которых ≥ 2 , затруднительно.

Однако не всегда при оценке точности воспроизведения сигнала в качестве критерия при выборе параметра используют квадратичную интегральную оценку. В некоторых случаях представляет интерес, например, возможно меньшее искажение фронта импульса или возможно меньшее время затухания переходного процесса, так называемое время выхода ИУ на квазистационарный режим работы (см. 2.3.5).

Рассмотрим вопрос о выборе критериев на примере воспроизведения единичной функции $1(t)$ ИУ, описываемым дифференциальным уравнением второго порядка. Выходной сигнал (2.269) такого устройства при $\beta < 1$ и воздействии $1(t)$ был получен в 2.3.4. Перепишем его еще раз для $S(0) = 1$:

$$y(t) = 1(t) \left[1 - \exp(-\omega_0 \beta t) \sin \left(\omega_0' t + \arctg \sqrt{\frac{1-\beta^2}{\beta^2}} \right) \right]. \quad (2.331)$$

На рис. 2.64 построены кривые выходного сигнала ИУ, имеющего частоту собственных колебаний $f_0 = 1000$ Гц, для различных степеней успокоения β и кривая входного сигнала $1(t)$.

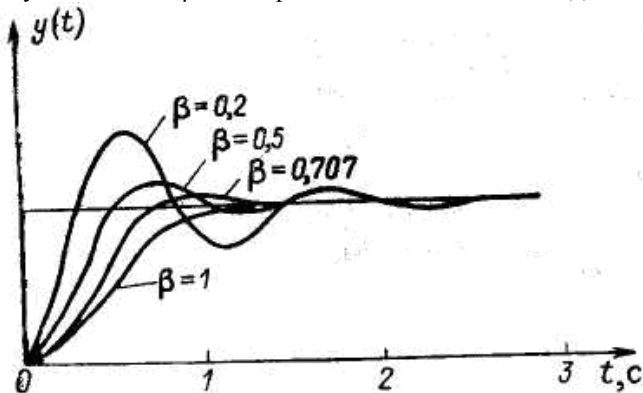


Рис. 2.64. Зависимость характера колебательного процесса от успокоения.

Из рассмотрения рисунка уже можно сделать заключение, что при $\beta = 0,7$ переходный процесс затухает быстрее всего, а при уменьшении β лучше воспроизводится фронт импульса и, вероятно, меньше среднее квадратическое значение динамической ошибки, так как при $\beta=0,7$ положительный переброс не компенсируется отрицательным.

Найдем степень успокоения, соответствующую минимуму времени успокоения, т. е. в качестве критерия выбора параметра β выберем минимум времени успокоения. **Временем успокоения**, как известно, называется время, по истечению которого можно принять, что переходный процесс закончился и система (ИУ) практически вышла в стационарный режим. Время успокоения, как уже говорилось, иногда называют и **временем выхода системы в квазистационарный режим**, т. е. режим, при котором отклонение $y(t)$ от $y(0)$ не превосходит некоторой заданной величины. Пусть задано

$$\delta = \frac{y(t) - y(0)}{y(0)} = 0,01. \quad (2.332)$$

Тогда для

$$\sin(\omega_0' t + \arctg \sqrt{1 - \beta^2/\beta^2}) = 1 \quad (2.333)$$

откуда

$$\delta = \frac{1}{\sqrt{1 - \beta^2}} \exp(-\beta \omega_0 t_{\text{усп}}), \quad (2.334)$$

Минимизируя $t_{\text{усп}}$, можно найти оптимальное значение β . Иметь малое время успокоения всегда желательно, но иногда другие критерии являются более важными. Если ставится задача наилучшего воспроизведения амплитуды импульса или его фронта, то в качестве критерия можно выбрать величину первого переброса, т. е.

$$\delta = \frac{y(t_1) - y(0)}{y(0)}, \quad (2.335)$$

где $y(t_1)$ —первый максимум $y(t)$, соответствующий моменту времени t_1 . Первый максимум $y(t)$ можно найти, продифференцировав (2.268) и приравняв производную нулю. Проделав это, получим

$$\exp(-\beta \omega_0 t_1) \sin \omega_0' t_1 = 0, \quad (2.336)$$

откуда находим время достижения максимума как

$$t_1 = \pi/\omega_0' = \pi/\omega_0 \sqrt{1 - \beta^2}$$

(второе значение t , равное нулю, тривиально).

При

$$t = t_1 \cos \omega'_0 t_1 = \cos \omega'_0 \pi / \omega'_0 = 1.$$

Тогда $y(t)$ в момент $t = t_1$ принимает значение

$$\begin{aligned} y(t_1) &= 1(t) [1 + \exp(-\beta \omega_0 t_1)] = \\ &= 1(t) \left[1 + \exp\left(-\frac{\beta \pi}{\sqrt{1-\beta^2}}\right) \right]. \end{aligned} \quad (2.337)$$

Подставляя (2.337) в (2.335) и учитывая, что $y(0) = 1$, получим

$$\delta = \exp\left(-\frac{\beta \pi}{\sqrt{1-\beta^2}}\right) \quad (2.338)$$

и

$$\beta = \frac{1}{\sqrt{\left(\frac{\pi}{\ln \delta}\right)^2 + 1}} \quad (2.339)$$

Минимизируя принятый критерий δ , можно выбрать β , соответствующее наименьшему искажению амплитуды сигнала; при этом заметнее искажается фронт импульса (см. рис. 2.64).

При минимизации квадратичного интегрального критерия для рассматриваемого ИУ при воздействии $1(t)$ импульса, т. е. минимизации

$$I = \int_0^{\infty} [\Delta_d(t)]^2 dt, \quad (2.340)$$

выражение для $\Delta_d(t)$ получается в виде

$$\begin{aligned} \Delta_d(t) &= y(t) - y(0) = -\exp(-\beta \omega_0 t) \sin\left(\omega'_0 t + \arctg \sqrt{\frac{1-\beta^2}{\beta^2}}\right) = \\ &= \exp(-\beta \omega_0 t) \left(\frac{\beta}{\sqrt{1-\beta^2}} \sin \omega'_0 t + \cos \omega'_0 t \right). \end{aligned} \quad (2.341)$$

Тогда

$$I = \int_0^{\infty} \exp(-2\beta \omega_0 t) \left(\frac{\beta}{\sqrt{1-\beta^2}} \sin \omega'_0 t + \cos \omega'_0 t \right)^2 dt. \quad (2.342)$$

После интегрирования

$$I = \frac{1}{\omega_0^2} \left(\frac{1}{4\beta} + \beta \right)^2. \quad (2.343)$$

Минимизируя I , получим $\beta = 0,5$.

В качестве примера найдем дисперсию динамической ошибки для ИУ, имеющего передаточную функцию

$$W(p) = \frac{1}{Tp + 1}, \quad (2.344)$$

при воздействии на него $\delta(t)$.

Так как спектральная плотность $\delta(t)$ равна единице, то

$$\sigma_{\Delta}^2 = D_{\Delta} = \int_0^{\infty} \left| \frac{1}{1 + j\omega T} \right|^2 d\omega = \int_0^{\infty} \frac{1}{1 + \omega^2 T^2} d\omega = \frac{\pi}{2T}. \quad (2.345)$$

2.3.8. Случайные сигналы и их характеристики

В практике часто приходится иметь дело с измеряемыми сигналами, являющимися случайными функциями времени. Так, случайными функциями времени являются параметры атмосферы для некоторой высоты, параметры атмосферы в герметичной кабине летательного аппарата и т. д. Случайными функциями времени часто оказываются и процессы, которые обычно принимаются за детерминированные. Например, задается постоянная скорость набора высоты, т. е. $H=vt$. Однако вследствие погрешностей систем управления и различных внешних влияний этот закон строго не выдерживается, и набор высоты будет происходить по законам, изображенным на рис. 2.65 кривыми 1, 2, 3.

Следовательно, измеряемые на борту параметры (вертикальная скорость и высота) являются случайными функциями времени. То же можно сказать и в отношении других параметров, которые поддерживаются на заданном уровне или должны изменяться по наперед заданному закону. Раз их надо «поддерживать», значит они как-то случайно измеряются во времени. Регулирующие устройства, выполняющие эту задачу, обеспечивают удержание параметра на заданном уровне для каждого момента времени с некоторой точностью. Измеряемый параметр попрежнему остается случайным процессом, но уже имеет другие характеристики.

Шумы и помехи, сопровождающие процессы измерения, также являются случайными функциями времени. Передаточные функции ИУ по помехам отличны от передаточных функций по измеряемому (полезному) сигналу. На выходе же ИУ получается один сигнал, определяемый как измеряемой величиной, так и помехой (шумами). Поэтому, естественно, встает задача выделения из него сигнала, зависящего от измеряемой величины.

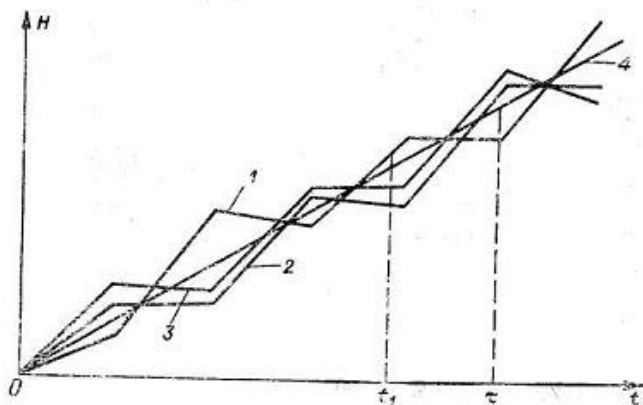


Рис. 2.65. Набор высоты летательным аппаратом как случайный процесс.

Случайный процесс отображается множеством непрерывных функций, каждая из которых представляет собой одну из возможных реализаций (рис. 2.66).

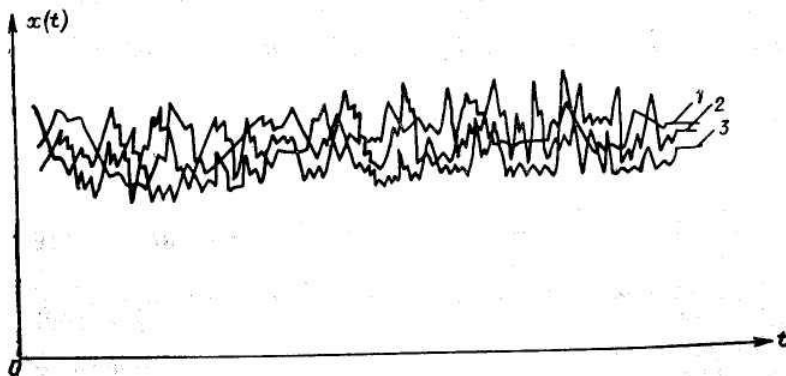


Рис. 2.66. Реализации случайного процесса.

Аналогично обозначениям случайных величин случайный процесс записывается как $X(t)$, а его реализации — $x_1(t)$, $x_2(t)$ и т. д. При измерении процесса имеют дело с одной из его реализаций. Задача же состоит в определении случайного процесса $X(t)$. Наиболее полно случайные процессы описываются законами распределения: одномерным, двумерным и т. д. Одномерный закон распределения

$f(x, t)$ характеризует только закон распределения случайной величины $X(t_1)$ для заданного, хотя и произвольного момента времени t_1 и не дает представления о зависимости $X(t)$ при различных t . Двумерный закон распределения $f(x_1, t_1; x_2, t_2)$ — это закон распределения системы двух случайных величин — $X(t_1)$ и $X(t_2)$. Очевидно что n -мерный закон еще более полно характеризует случайный процесс $X(t)$. Однако использование закона распределения более высокого порядка, чем второй, затруднительно. Для практических целей обычно бывает достаточно знать характеристики случайного процесса — моменты первого и второго порядков. Ограничимся рассмотрением стационарных случайных процессов, для которых моменты первого и второго порядков не зависят от времени, т. е. математическое ожидание и дисперсия являются постоянными числовыми характеристиками, а корреляционная функция зависит только от интервала времени — интервала корреляции. **Случайный процесс называется стационарным, если плотности распределения вероятностей для моментов времени t_1, t_2, \dots, t_n взаимно тождественны.** Большинство стационарных случайных процессов обладает свойством **эргодичности**: *среднее по времени на достаточно большом участке наблюдения приближенно равно среднему по множеству наблюдений*, т. е.

$$\tilde{x} = \int_{-\infty}^{\infty} x f(x) dx = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t) dt = \bar{x}. \quad (2.346)$$

Это значит, что отдельная реализация стационарного случайного процесса на бесконечном интервале времени полностью определяет случайный процесс $X(t)$, все бесчисленное множество его реализации. Таким образом, для стационарного процесса, обладающего свойством эргодичности, математическое ожидание

$$m_x = \tilde{x} = \bar{x}, \quad (2.347)$$

момент второго порядка

$$a_2 = \int_{-\infty}^{\infty} x^2 f(x) dx = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x^2 dt = \bar{x}^2 = \tilde{x}^2, \quad (2.348)$$

т. е. равен среднему квадрату случайной величины $X(t)$. Дисперсия $X(t)$ — второй центральный момент — равна

$$D_x = \int_{-\infty}^{\infty} (x - \bar{x})^2 f(x, t) dx = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T (x - \bar{x})^2 dt. \quad (2.349)$$

Наиболее полно случайные процессы характеризуются корреляционной функцией и энергетическим спектром.

Корреляционная функция — начальный корреляционный момент двух значений случайного процесса $x(t)$ и $x_1(t_1)$ — определяется как

$$K_x(t, t_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t) \cdot x(t_1) \cdot f(x, t; x_1, t_1) dx dx_1, \quad (2.350)$$

или

$$K_x(t, t_1) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) \cdot x_1(t_1) dt. \quad (2.351)$$

Центральный корреляционный момент

$$\begin{aligned} \overset{\circ}{K}_x(t, t_1) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x(t) - \bar{x}] [x_1(t_1) - \bar{x}] dt = \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x(t) - \bar{x}] [x(t - \tau) - \bar{x}] dt, \end{aligned} \quad (2.352)$$

где $\tau = t - t_1$ — интервал корреляции.

Корреляционная функция эргодического стационарного процесса обладает следующими свойствами:

- 1) при $\tau=0$ $K_x(0) = [x(t)]^2$; $\overset{\circ}{K}_x(0) = D_x(t)$;
- 2) при $\tau=\infty$ $K_x(\infty) = (\tilde{x}^2) = (\bar{x})^2$; $\overset{\circ}{K}_x(\infty) = 0$;
- 3) является четной функцией и имеет максимум при $\tau=0$.

Корреляционные функции могут применяться для описания свойств не только случайных, но и детерминированных процессов. Например, корреляционная функция постоянной величины $x(t) = X_0 = \text{const}$ имеет вид

$$K_x(\tau) = X_0^2. \quad (2.353)$$

Для периодической функции $x(t) = X_{1m} \sin(\omega_1 t + \varphi_1)$

$$K_x(\tau) = \frac{X_{1m}^2}{2} \cos \omega_1 \tau. \quad (2.354)$$

Энергетический спектр характеризует среднюю мощность процесса на интервале наблюдения. Он определяется как

$$G(\omega) = \lim_{T \rightarrow \infty} \frac{1}{2T} |S_x(j\omega)|^2. \quad (2.355)$$

Используя пару преобразований Фурье [см. формулы (2.202) и (2.204)], можно показать, что для стационарного процесса

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} |S_x(j\omega)|^2 d\omega = \int_{-\infty}^{\infty} [x(t)]^2 dt. \quad (2.356)$$

Так как величина, стоящая в правой части (2.356), пропорциональна энергии процесса, то левая часть называется энергетической формой интеграла Фурье. Если поделить (2.356) на $2T$ и взять предел при $T \rightarrow \infty$, то с учетом (2.355) получим

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) d\omega = \overline{x^2}. \quad (2.357)$$

Корреляционные функции и энергетические спектры связаны между собой взаимными преобразованиями Фурье:

$$K_x(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) \exp(j\omega\tau) d\omega;$$

$$G_x(\omega) = \int_{-\infty}^{\infty} K_x(\tau) \exp(-j\omega\tau) dt. \quad (2.358)$$

Любой стационарный случайный процесс $X(t)$ выражается интегральным каноническим представлением:

$$X(t) = m_x + \int_{-\infty}^{\infty} V(\omega) \exp(j\omega t) d\omega, \quad (2.359)$$

где $V(\omega)$ —случайная (центрированная) функция с некоррелированными значениями при различных ω , интенсивность которой равна энергетическому спектру $G_v(\omega)$ случайной величины $X(t)$. Эта функция обычно называется «белым шумом» и широко используется при исследовании систем автоматического управления и регулирования. Энергетический спектр белого шума имеет одинаковое значение во всем бесконечном диапазоне частот, т. е.

$$G_v(\omega) = G_0 = \text{const}, \quad (2.380)$$

и, следовательно, корреляционная функция, на основании (2.358)

$$K_V(\tau) = \frac{1}{\pi} \int_0^{\infty} G_0 \exp(j\omega\tau) d\omega = G_0 \delta(t), \quad (2.381)$$

так как $\exp(j\omega t)$ —спектральная плотность δ -функции, а

$$\frac{1}{\pi} \int_0^{\infty} \exp(j\omega t) d\omega = \delta(t). \quad (2.382)$$

Таким образом, для белого шума характерно отсутствие корреляционной связи между значениями случайной величины при всех значениях t , кроме $t = 0$ (рис. 2.67, а).

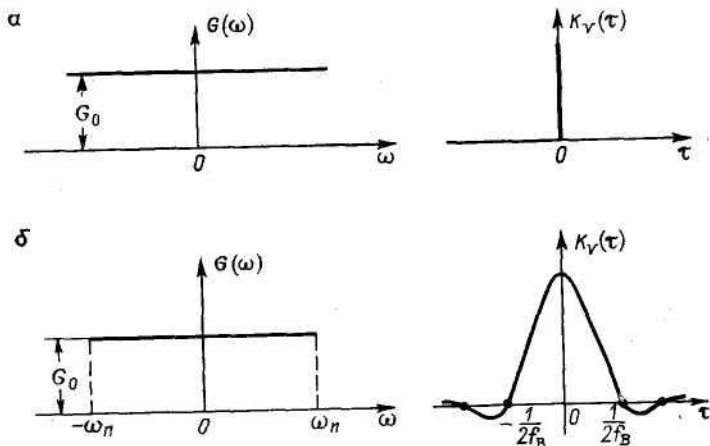


Рис. 2.67. Энергетический спектр и корреляционная функция белого шума.

а — теоретические; б — для реальных процессов.

Это физически нереальный процесс, так как ему соответствует бесконечно большая дисперсия [см. (2.349)]

$$D_V = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\infty}^{\infty} [V(t)]^2 dt \quad (2.363)$$

и, следовательно, бесконечно большая мощность. Для характеристики физически реальных процессов используется белый шум с ограниченным энергетическим спектром, как это показано на рис. 2.67, б. Диапазон частот спектра

$$\Delta f = \frac{\Delta\omega}{2\pi} = 2f_B. \quad (2.364)$$

Так как корреляционная функция белого шума с ограниченным спектром

$$K_V(\tau) = \frac{G_0}{\pi\tau} \sin \omega_B \tau, \quad (2.365)$$

то

$$D_V = K_V(0) = \frac{G_0 2\omega_B}{2\pi} = \frac{G_0 \omega_B}{\pi}. \quad (2.366)$$

Возможны и другие представления случайных процессов

2.3.9. Реакция измерительных устройств на случайные возмущения. Характеристики точности

Случайное возмущение, действующее на входе прибора, как уже отмечалось, представляет собой смесь полезного сигнала и помехи. В общем случае это некоррелированные сигналы. Поэтому вначале рассмотрим вопрос о возможности описания выходного сигнала стационарного ИУ при действии на его вход случайного измеряемого процесса. В общем виде, так же как это имело место при детерминированных воздействиях,

$$y(t) = W(p) \cdot x(t). \quad (2.367)$$

Однако передаточная функция $W(p) = y(t)/x(t)$ для случайного процесса на входе непосредственно не известна и может быть определена лишь косвенным путем. Соответственно выходной сигнал для случая произвольного случайного воздействия может быть определен только по некоторым известным характеристикам сигнала и устройства. Если входной сигнал представлен в интегральном каноническом виде

$$x(t) = m_x + \int_{-\infty}^{\infty} V(\omega) \exp(j\omega t) d\omega, \quad (2.368)$$

то выходной сигнал, в силу принципа суперпозиции в интегральной форме для линейного ИУ, должен иметь аналогичное интегральное каноническое представление выходной переменной. Однако практическое определение его связано с рядом трудностей и может быть выполнено только численными или графическими методами. Аналитически можно найти корреляционную функцию выходного сигнала и его дисперсию, которые являются характеристиками выходного сигнала. Так как [см. (2.357)]

$$K_y(\tau) = K_y(t, t_1) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{G}_y(\omega) \exp[j\omega(t - t_1)] d\omega, \quad (2.369)$$

то для ее определения надо найти модуль спектральной функции выходного сигнала $G_y(\omega)$. Спектральная плотность выходного сигнала

$$S_y(j\omega) = S(j\omega) \cdot S_x(j\omega), \quad (2.370)$$

где $S(j\omega)$ —комплексная чувствительность ИУ.

Энергетический спектр входного сигнала

$$G_x(\omega) = \lim_{T \rightarrow \infty} \frac{1}{2T} |S_x(j\omega)|^2.$$

Тогда энергетический спектр выходного сигнала

$$\begin{aligned} G_y(\omega) &= \lim_{T \rightarrow \infty} \frac{1}{2T} |S_y(j\omega)|^2 = \lim_{T \rightarrow \infty} \frac{1}{2T} |S(j\omega) \cdot S_x(j\omega)|^2 = \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} |S_x(j\omega)|^2 \cdot |S(j\omega)|^2 = G_x(\omega) \cdot |S(j\omega)|^2. \end{aligned} \quad (2.371)$$

Таким образом, энергетический спектр выходного сигнала равен произведению энергетического спектра $x(t)$ на квадрат модуля комплексной чувствительности измерительного устройства.

Дисперсия выходного сигнала

$$D = K_y(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_y(\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_x(\omega) |S(j\omega)|^2 d\omega, \quad (2.372)$$

Ошибка воспроизведения измеряемого сигнала

$$\Delta_n(t) = y(t) - x(t), \quad (2.373)$$

где $x(t)$ и $y(t)$ — реализация входного сигнала и соответствующая ей реализация выходного сигнала; $\Delta_n(t)$ —также случайный процесс, характеризуемый спектральной плотностью, корреляционной функцией и дисперсией. Если $G_\Delta(\omega)$ —энергетический спектр ошибки, то

$$K_\Delta(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_\Delta(\omega) \exp(j\omega\tau) d\omega \quad (2.374)$$

есть корреляционная функция ошибки, а

$$D_\Delta = K_\Delta(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_\Delta(\omega) d\omega \quad (2.375)$$

— ее дисперсия.

Дисперсия, средний квадрат ошибки по времени $\overline{\Delta_0^2}$ является весьма распространенной оценкой качества воспроизведения сигнала. Энергетический спектр ошибки как случайной величины связан с энергетическим спектром сигнала выражением, аналогичным (2.371), и может быть записан как

$$G_{\Delta}(\omega) = |S_{\Delta}(j\omega)|^2 G_x(\omega), \quad (2.376)$$

где $S_{\Delta}(j\omega) = S(j\omega) - 1$ — комплексная чувствительность ИУ по ошибке.

Таким образом, ошибка воспроизведения случайного стационарного процесса линейным ИУ находится совершенно аналогично ошибке воспроизведения детерминированного сигнала. Аналитическое определение выходного сигнала может быть сделано, если экспериментально определить корреляционные функции входного $K_x(\tau)$ и выходного $K_y(\tau)$ сигналов. Корреляционную функцию выходного сигнала часто называют взаимной корреляционной функцией между входным и выходным сигналами. Действительно, если в (2.351) значение функции $x_1(t_1) = -x(t + \tau)$ принять за $y(t + \tau)$ и выразить ее через $x(t + \tau)$ и, например, весовую функцию, то

$$y(t) = \int_{-\infty}^{\infty} x(t + \tau - \theta) \cdot g(\theta) d\theta \quad (2.377)$$

и

$$\begin{aligned} K_{xy}(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) \cdot y(t + \tau) dt = \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) \int_0^{\infty} x(t + \tau - \theta) \cdot g(\theta) d\theta dt = \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^{\infty} \int_0^T x(t) \cdot x(t + \tau - \theta) \cdot g(\theta) d\theta dt = \\ &= \int_0^{\infty} K_x(\tau - \theta) \cdot g(\theta) d\theta, \end{aligned} \quad (2.378)$$

где $K_x(\tau - \theta)$ — корреляционная функция входного сигнала.

Из (2.378) видно, что непосредственно определить комплексную чувствительность по $K_{xy}(\tau)$ и $K_x(t - \theta)$ нельзя, можно лишь, подав на вход белый шум и определив $K_{xy}(\tau)$, найти весовую функцию $g(\theta)$. Действительно, так как корреляционная функция белого шума есть δ -функция, то

$$K_{xy}(\tau) = \int_0^{\infty} \delta(\tau - \theta) \cdot g(\theta) d\theta = g(\tau). \quad (2.379)$$

Для нахождения $S(j\omega)$ нужно по экспериментально снятым $K_{xy}(\tau)$ и $K_x(\tau-\Theta)$ найти спектральные плотности $S_{xy}(j\omega)$ и $S_x(j\omega)$. Тогда комплексная чувствительность

$$S(j\omega) = \frac{S_{xy}(j\omega)}{S_x(j\omega)}. \quad (2.390)$$

Следует отметить, что функция взаимной корреляции является более полной оценкой точности воспроизведения сигнала, чем дисперсия ошибки (2.375), так как она учитывает временной сдвиг Θ , возникающий вследствие задержки (чистой задержки) выходного сигнала по отношению к входному.

Если на вход ИУ действует сигнал $x(t)$ и помеха $u(t)$, т. е. сигнал

$$z(t) = x(t) + u(t),$$

то корреляционная функция сигнала $z(t)$

$$\begin{aligned} K_z(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x(t) + u(t)] [x(t + \tau) + u(t + \tau)] dt = \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) \cdot x(t + \tau) dt + \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T u(t) \cdot u(t + \tau) dt + \\ &+ \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t + \tau) \cdot u(t) dt + \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) \cdot u(t + \tau) dt = \\ &= K_x(\tau) + K_u(\tau) + K_{xu}(\tau) + K_{ux}(\tau). \end{aligned} \quad (2.381)$$

При достаточно большом времени усреднения $K_{xu}(\tau)$ и $K_{ux}(\tau)$ равны нулю, так как $x(t)$ и $u(t)$ статистически независимы. Тогда

$$K_z(\tau) = K_x(\tau) + K_u(\tau).$$

Так как помеха обычно имеет достаточно широкий спектр частот и быстро затухающую корреляционную функцию, то с увеличением τ $K_u(\tau)$ стремится к нулю и наблюдаемая функция $K_z(\tau) \rightarrow K_x(\tau)$. Таким образом, посредством коррелятора можно выделить сигнал на фоне помехи или во всяком случае повысить отношение «сигнал/шум».

2.3.10. Определение динамических свойств измерительных устройств с помощью моделей-аналогов

Уравнение (2.378) называется основным уравнением идентификации. Оно позволяет по корреляционным функциям входного и

выходного сигналов определить динамические свойства ИУ в режиме функционирования, так как, во-первых, корреляционные функции дают наиболее полное представление о случайных процессах и, во-вторых, могут быть выражены аналитически достаточно строго. Уравнением идентификации является и уравнение (2.367). Для определения передаточной функции при этом необходимо иметь точное аналитическое описание входного и выходного сигналов, что, как уже неоднократно отмечалось, сделать весьма трудно, если сигнал на входе — произвольный случайный процесс. Именно это обстоятельство и ограничивает возможности его использования.

Идентификация ИУ его математической моделью—это процесс определения математической модели ИУ с целью исследования его динамических характеристик с использованием реально действующих на его входах сигналов и реакций ИУ на эти сигналы.

В 2.3.9 была рассмотрена идентификация ИУ математической моделью, выраженной в аналитической форме. Такой метод решения задачи требует аппроксимации корреляционных функций, что связано с большими вычислительными трудностями. Поэтому независимо от точности метода решение уравнения идентификации получится приближенным.

В связи с этим в ряде случаев ***целесообразнее использовать численные методы решения, заменяя уравнение идентификации системой алгебраических уравнений.*** Такая система достаточно просто может быть решена с помощью вычислительных машин. Еще более перспективным методом является решение задачи идентификации ИУ в динамическом режиме с помощью электронных моделей-аналогов, изоморфных ИУ по поведению. Метод идентификации с помощью ***моделей-аналогов*** получил широкое распространение, что может быть объяснено попыткой исключить следующие недостатки аналитических и численных методов решения:

- 1) несоблюдение для реальных условий основных ограничений, принятых при выводе уравнений идентификации;
- 2) несоответствие корреляционных функций, определяемых в процессе функционирования диагностируемой системы, их истинным значениям и недооценка (в некоторых системах) коррелированности внутренних помех и внешних возмущений;
- 3) невозможность учета явных и неявных обратных связей, имеющих в реальной измерительной системе, и т. д.

Использование моделей-аналогов в ряде случаев может быть существенно расширено в особенности для проведения многократного периодически повторяющегося экспресс-анализа.

Модели могут быть реализованы на аналоговых элементах (на АВМ) и на ЦВМ. **В основе аналоговых моделей лежит математическая теория подобия.** Их достоинством является возможность проведения экспресс-анализа исследуемой системы в любой момент времени протекания процесса в непрерывном взаимодействии всех его параметров. **Дискретные моделирующие устройства** имеют большую степень общности; операции, проводимые с их помощью, сводятся к небольшому выбору элементарных действий.

Сформировались два метода математического описания динамических систем с целью определения их характеристик:

- 1) аппроксимация весовой функции системы суммой конечного числа членов ряда, составленного из динамических элементов некоторой ортогональной системы функций;
- 2) определение статистической модели динамического объекта методом множественной регрессии.

1. Аппроксимация весовой функции системы суммой конечного числа членов ряда, составленного из динамических элементов некоторой ортогональной системы функций, является наиболее разработанным способом построения модели-аналога.,

В общем виде уравнение аппроксимации может быть записано в виде

$$g_m(t - \tau) = \sum_{k=1}^n C_k \varphi_k, \quad (2.382)$$

где φ_k — ортогональная система функций, например, полиномы Лагерра или Каутца.

Реализацию модели целесообразно выполнить на ЭВМ. Настройка модели производится по минимуму среднего квадратичного отклонения

$$\sigma^2 = \int_0^{\infty} [g(t - \tau) - g_m(t - \tau)]^2 d\tau \quad (2.383)$$

при подаче на вход наряду с измеряемым случайным процессом пробного сигнала в виде двоичного белого шума. При этом составляющая выхода, обусловленная действием измеряемого сигнала, должна рассматриваться как помеха, характеризующаяся энергетическим спектром $G_x(\omega)$.

2. Метод множественной регрессии хорошо разработан для моделирования статических ИУ. Сигнал на входе в общем случае— это

многомерный случайный процесс, в состав которого, кроме полезного измеряемого сигнала, входят помехи и внешние факторы. Так, при измерении влажности на вход системы действуют температура и давление среды, в которой расположен датчик, и т. п. Аналитическое исследование динамики систем при многомерном входном сигнале сложно, требует больших вычислительных работ и не обеспечивает должной точности (передаточные функции по составляющим входного сигнала разные). При использовании для анализа ИУ в многофакторном пространстве моделей-аналогов, рассмотренных выше, возникает ряд трудностей, связанных с настройкой модели.

Статическое ИУ, на которое воздействует множество случайных возмущений, обычно представляется в виде

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \Delta, \quad (2.384)$$

где x_i — входные случайные стационарные процессы, статистически независимые; Δ — помеха, приведенная к выходу; β_i — коэффициенты регрессии.

Оценки β_i для (2.384) легко определяются, если ИУ не содержит динамических каналов. В этом случае уравнение модели ИУ

$$y_m = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_n x_n. \quad (2.385)$$

Неоднозначность полученной оценки (модели) нарушается действием помехи Δ . Если же среди каналов ИУ есть динамические звенья, то значение y определяется не только x , но и всей предысторией входных переменных. При неблагоприятных соотношениях между статистическими характеристиками входного сигнала $x(t)$ и параметрами динамического канала неопределенность, вносимая динамикой в (2.385), будет столь значительной, что (2.385) не может рассматриваться не только как динамическая модель, но даже как статическая. Если весовые функции каналов ИУ $g_k(\tau)$, то для линейного ИУ можно написать

$$y(\tau) = \sum_{k=1}^n \int_0^{\infty} g_k(t - \tau) x_k(\tau) d\tau. \quad (2.386)$$

При определении динамической модели ИУ можно в каждый канал ввести звено чистого запаздывания. Тогда регрессионная модель (динамическая) представляется уравнением

$$y_m(\tau) = \sum_{k=1}^n x_k(t - \tau_k) \beta_k, \quad (2.387)$$

где τ_k — постоянная времени k -го канала. Функция $y(\tau)$ и $y_m(\tau)$ отличаются на величину ошибки аппроксимации динамики $\Delta_d(\tau)$, т. е.

$$y(\tau) = y_m(\tau) + \Delta_d(\tau) + \Delta. \quad (2.388)$$

Точность аппроксимации динамики ИУ зависит от оптимальности выбора постоянных времени каналов. Критерием оптимальности выбора постоянных времени является минимум

$$D|\Delta_d(\tau) + \Delta| = \sigma^2.$$

Для создания динамических моделей можно предложить метод подбора коэффициентов дифференциального уравнения. При аппроксимации импульсной переходной функции ИУ в многофакторном пространстве суммой конечного числа членов ряда, составленного из динамических элементов некоторой ортогональной системы функций, число элементов получается настолько большим, что точность настройки модели оказывается очень низкой. Следует принять во внимание также, что *исследование ИУ*, как правило, *предполагает не только определение оценки погрешности, но и выяснение причин, определяющих погрешность*. С этой точки зрения модели, построенные на динамических элементах ортогональной системы функций, ничего не дают. Значительно большую информацию о работе ИУ можно получить, представив его модель в виде дифференциального уравнения, так как связь между коэффициентами уравнения и параметрами ИУ в большинстве случаев известна. Порядок уравнения модели может быть принят равным двум и в некоторых случаях даже единице. Такие рекомендации основываются на известном положении, что линейная система сколь угодно высокого порядка в переходном процессе при любом возмущении может быть описана уравнением третьего или второго порядка. Утверждение не является безусловным, однако для сравнительно небольших скоростей изменения входного воздействия оно может быть принято.

В многофакторном пространстве коэффициенты уравнения модели являются случайными функциями факторов, в том числе измеряемой величины. Обычно модель ИУ выполняют в виде дифференциального уравнения с постоянными коэффициентами, учет же влияния внешних факторов (чаще всего температуры) осуществляется путем выбора некоторых средних значений коэффициентов уравнения, обеспечивающих минимизацию принятого при настройке модели критерия.

Рассмотрим модель ИУ с переменными коэффициентами, обеспечивающую с помощью небольших регулировок получение результатов, аналогичных результатам, снимаемым с ИУ в многофакторном пространстве. Примем в качестве математической модели ИУ дифференциальное уравнение второго порядка

$$a_2 \frac{d^2 y(t)}{dt^2} + a_1 \frac{dy(t)}{dt} + a_0 y(t) = x_1(t), \quad (2.389)$$

где a_0, a_1, a_2 — коэффициенты уравнения, зависящие от многомерного случайного процесса (от влияющих факторов, изменяющихся случайным образом).

Настройка модели, определение ее коэффициентов производится в процессе функционирования ИУ при подаче на его вход типового воздействия. Выбор вида входного воздействия (измеряемого процесса) зависит от возможностей его реализации. Предпочтительно иметь на входе импульсную или единичную функцию либо синусоидальную.

Эксперимент планируется таким образом, чтобы в результате можно было получить зависимость коэффициентов уравнения модели от всех влияющих факторов для заданного входного воздействия (при различных его уровнях). Уровни входного воздействия и влияющих факторов задаются дискретно. По табл. 3.2 для примера приведен план эксперимента для случая трех влияющих факторов. Измеряемый процесс $x_1(t)$ задается в виде типового воздействия различных уровней: X_{11}, X_{12}, X_{13} . Влияющие факторы заданного уровня в соответствии с планом эксперимента устанавливаются до подачи измеряемого процесса (первый индекс у $X(t)$ — номер фактора, в том числе измеряемого процесса; второй — указывает уровень фактора).

Для каждого эксперимента фиксируется (записывается) выходной сигнал ИУ. Индексы у $Y(t)$ обозначают уровни факторов, начиная с первого. После составления плана эксперимента производится настройка модели для каждого из экспериментов таблицы 3.2 путем изменения коэффициентов модели при минимизации

$$\sigma^2 = \frac{1}{m} \sum_{r=1}^m [y(r) - y_m(r)]^2,$$

где $y_m(t)$ — выходной сигнал модели; $r = 1, 2, 3, \dots$ при интервале между точками Δt ; $y(r), y_m(r)$ — значения ординат выходного сигнала ИУ и модели в фиксированных точках.

Таблица 3.2

План эксперимента

№ экспери- мента	Измеряе- мый про- цесс			Влияющие факторы													Y (t)
	X ₁₁	X ₁₂	X ₁₃	X ₂₀	X ₂₁	X ₂₂	X ₂₃	X ₃	X ₃₁	X ₃₂	X ₃₄	X ₄₀	X ₁₁	X ₁₂	X ₁₃		
1	+			+				-					+			Y ₁₀₀₀	
2		+		+				+					+			Y ₂₀₀₀	
3			+	+				+					+			Y ₃₀₀₀	
4	+				+			+					+			Y ₁₁₀₀	
5		+			+			+					+			Y ₂₁₀₀	
6			+		+			+					+			Y ₃₁₀₀	
7	+					+		+					+			Y ₁₂₀₀	
8		+				+		+					+			Y ₂₂₀₀	
9			+			+		+					+			Y ₃₂₀₀	
10	+						+	+					+			Y ₁₃₀₀	
11		+					+	+					+			Y ₂₃₀₀	
12			+				+	+					+			Y ₃₃₀₀	
13	+			+					+				+			Y ₁₀₁₀	
14		+		+					+				+			Y ₂₀₁₀	
15			+	+					+				+			Y ₃₀₁₀	
16	+			+						+			+			Y ₁₀₂₀	
17		+		+						+			+			Y ₂₀₂₀	
18			+	+						+			+			Y ₃₀₂₀	
19	+			+							+		+			Y ₁₀₃₀	
20		+		+							+		+			Y ₂₀₃₀	
21			+	+							+		+			Y ₃₀₃₀	
22	+			+				+						+		Y ₁₀₀₁	
23		+		+				+						+		Y ₂₀₀₁	
24			+	+				+						+		Y ₃₀₀₁	
25	+			+				+							+	Y ₁₀₀₂	
26		+		+				+							+	Y ₂₀₀₂	
27			+	+				+							+	Y ₃₀₀₂	
28	+			+				+								Y ₁₀₀₃	
29		+		+				+								Y ₂₀₀₃	
30			+	+				+								Y ₃₀₀₃	

После того как настройка модели закончена (для данного эксперимента), записываются значения коэффициентов уравнения в табл. 3.3. В последнем столбце таблицы записывается минимальное значение среднего квадратичного отклонения, при котором была закончена настройка модели. Следует заметить, что число экспериментов может быть значительно уменьшено, если известно, что ИУ является линейной системой.

Таблица 3.3

Результат эксперимента

№ экспери- мента	Уровни факторов				a_2	a_1	a_0	σ_{\min}^2
	$x_1(t)$	$x_2(t)$	$x_3(t)$	$x_4(t)$				
1	X_{11}	X_{20}	X_{30}	X_{40}				
2	X_{12}	X_{20}	X_{30}	X_{40}				
3	и т. д.							

По данным табл. 3.3 можно построить серию кривых, отражающих зависимость коэффициентов уравнения от влияющих факторов для различных уровней измеряемого процесса.

Используя зависимости коэффициентов уравнения от влияющих факторов, полученные в результате экспериментального исследования реального устройства, можно достаточно просто найти модель-аналог для различных уровней измеряемого сигнала и различных сочетаний влияющих факторов.

3. Параметры, факторы, модели объектов исследования

3.1. Параметры оптимизации

При планировании эксперимента очень важно определить параметр, который нужно оптимизировать. Сделать это довольно затруднительно. Цель исследования должна быть сформулирована очень четко и допускать количественную оценку. Будем называть *характеристику цели*, заданную количественно, **параметром оптимизации**. **Параметр оптимизации является реакцией (откликом) на воздействие факторов, которые определяют поведение выбранной системы.** Реакция объекта многогранна, многоаспектна. Выбор того аспекта, который представляет наибольший интерес, как раз и задается целью исследования.

При традиционном нематематическом подходе исследователь стремится как-то учесть разные аспекты, взвесить их и принять *согласованное* решение о том, какой опыт *лучше*. Однако разные экспериментаторы проведут сравнение опытов неодинаково. Различия - одно из проявлений таланта исследователя или его бездарности.

Прежде чем сформулировать требования к параметрам оптимизации и рекомендации по их выбору, рассмотрим различные виды параметров.

3.1.1. Виды параметров оптимизации

В зависимости от объекта и цели исследования параметры оптимизации могут быть весьма разнообразными. Чтобы ориентироваться в этом многообразии, представим классификацию параметров (рис. 3.1), предложенную Ю.П.Адлером.

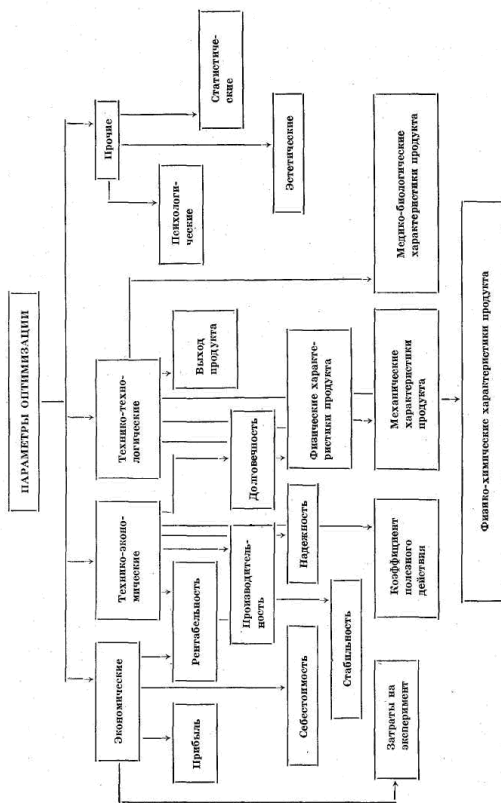


Рис. 3.1. Классификация параметров оптимизации

В основу любой классификации должна быть положена концепция, объясняющая **классификационные признаки**. Классификация представляет собой многоступенчатое, как правило, иерархическое, разветвленное деление логического объема понятия. В результате создается система соподчиненных понятий. Такой принцип классификации был взят за основу предложенной классификации параметров оптимизации.

Здесь не приводится полная и детальная классификации. Ставится задача — построить такую условную схему, которая включала бы ряд практически важных случаев и помогала экспериментатору ориентироваться в реальных ситуациях.

Реальные ситуации, как правило, сложны. Они часто требуют одновременного учета нескольких, иногда очень многих, параметров. В принципе каждый объект может характеризоваться сразу всей совокупностью параметров, приведенных на рис. 3.1, или любым подмножеством из этой совокупности. Движение к оптимуму возможно, если выбран один-единственный параметр оптимизации. Тогда прочие характеристики процесса уже не выступают в качестве параметров оптимизации, а служат **ограничениями**. Другой путь — построение обобщенного параметра оптимизации как некоторой функции от множества исходных .

Проанализируем предложенную схему классификации параметров. Экономические параметры оптимизации, такие, как прибыль, себестоимость и рентабельность, обычно используются при исследовании действующих промышленных объектов, тогда как затраты на эксперимент имеет смысл оценивать в любых исследованиях, в том числе и лабораторных. Если цена опытов одинакова, затраты на эксперимент пропорциональны числу опытов, которые необходимо поставить для решения данной задачи. Это в значительной мере определяет выбор плана эксперимента.

Среди технико-экономических параметров наибольшее распространение имеет производительность. Такие параметры, как долговечность, надежность и стабильность, связаны с длительными наблюдениями. Уже накоплен определенный опыт их использования при изучении дорогостоящих ответственных объектов, например радиоэлектронной аппаратуры.

Почти во всех исследованиях приходится учитывать количество и качество получаемого продукта. Как меру количества продукта используют выход, например, процент выхода химической реакции, выход годных изделий.

Показатели качества чрезвычайно разнообразны. В предложенной схеме они сгруппированы по видам свойств. Характеристики количества и качества продукта образуют группу технико-технологических параметров.

Под рубрикой «прочие» сгруппированы различные параметры, которые встречаются реже, но это не умаляет их значимость. Сюда попали статистические параметры, используемые для улучшения характеристик случайных величин или случайных функций. В качестве примеров назовем задачи на минимизацию дисперсии случайной величины, на уменьшение числа выбросов случайного процесса за фиксированный уровень и т. д. Последняя задача возникает, в частности, при выборе оптимальных настроек автоматических регуляторов или при улучшении свойств нитей (проволока, пряжа, искусственное волокно и др.).

С ростом сложности объекта возрастает роль психологических аспектов взаимодействия человека или животного с объектом. Так, при выборе оптимальной организации рабочего места оператора параметром оптимизации может служить число ошибочных действий в различных возможных ситуациях. Сюда относятся задачи выработки условных рефлексов типа задачи «крысы в лабиринте».

При решении задачи технической эстетики или сравнении произведений искусства возникает потребность в эстетических параметрах. Они основаны на ранговом подходе, который будет рассмотрен ниже.

Таковы некоторые виды параметров оптимизации.

Приведем следующий пример, взятый из литературных источников.

Пример 1. Во время второй мировой войны несколько сот английских торговых судов на Средиземном море были вооружены зенитными орудиями для защиты от вражеских бомбардировщиков. Поскольку это мероприятие было достаточно дорогим (требовалось иметь на каждом судне боевую команду), через несколько месяцев решили оценить его эффективность. Какой из параметров оптимизации более подходит для этой цели?

Число сбитых самолетов.

Потери в судах, оснащенных орудиями, по сравнению с судами без орудий.

Если Вы считаете, что эффективность установления орудий на торговые суда можно оценить числом сбитых самолетов, то Вы вряд ли смогли бы занять пост командующего английским флотом на Средиземном море. Выбранный Вами параметр оптимизации оценивает эффективность уничтожения самолетов. В то же время ясно,

что значения параметра оптимизации в этом случае будут низкими, так как существуют куда более эффективные средства для этой цели (авиация, боевой флот), чем зенитные орудия на торговых судах.

Если же Вы полагаете, что эффективность установки орудий на торговые суда можно оценить сопоставлением потерь в судах, оснащенных орудиями, с потерями в судах без орудий, то это разумный выбор параметра оптимизации, потому что основной задачей при установке орудий была защита судов. Самолеты вынуждены были теперь использовать противозенитные маневры и бомбометание с большой высоты, что уменьшало потери. Из числа атакованных самолетами торговых судов с зенитными орудиями было потоплено 10% судов, а потери в судах без орудий составили 25%. Затраты на установку орудий и содержание боевых расчетов окупались очень быстро.

3.1.2. Требования к параметру оптимизации

Параметр оптимизации — это признак, по которому следует оптимизировать процесс. Он должен быть количественным, задаваться числом. Мы должны уметь его измерять при любой возможной комбинации выбранных уровней факторов. Множество значений, которые может принимать параметр оптимизации, будем называть *областью его определения*. Области определения могут быть *непрерывными и дискретными, ограниченными и неограниченными*. Например, выход реакции — это параметр оптимизации с непрерывной ограниченной областью определения. Он может изменяться в интервале от 0 до 100%. Число бракованных изделий, число зерен на шлифе сплава, число кровяных телец в пробе крови — это примеры параметров с дискретной областью определения, ограниченной снизу.

Для измерения параметра оптимизации необходимо располагать требуемым прибором. Если нет способа количественного измерения параметра, то приходится воспользоваться способом, называемым *ранжированием (ранговым подходом)*. При этом параметрам оптимизации присваиваются оценки — ранги по заранее выбранной шкале: двухбалльной, пятибалльной и т. д. Ранговый параметр имеет дискретную ограниченную область определения. В простейшем случае область содержит два значения (да, нет; хорошо, плохо). Это может соответствовать, например, годной продукции и браку.

Ранг — это количественная оценка параметра оптимизации, но она носит условный (субъективный) характер. Мы ставим в соответствие качественному признаку некоторое число — ранг.

Для каждого физически измеряемого параметра оптимизации можно построить ранговый аналог. Потребность в построении такого аналога возникает, если имеющиеся в распоряжении исследователя численные характеристики неточны или неизвестен способ построения удовлетворительных численных оценок. При прочих равных условиях всегда нужно отдавать предпочтение физическому измерению, так как ранговый подход менее чувствителен и с его помощью трудно изучать тонкие эффекты.

Пример 2. Ваша жена решила испечь яблочный пирог по новому рецепту (аналогичный пример рассмотрен в литературе). Вам, конечно, трудно остаться в стороне, и вы предлагаете ей свои услуги по оптимизации этого процесса. Цель процесса — получение вкусного пирога, но такая формулировка цели еще не дает возможности приступить к оптимизации: необходимо выбрать количественный критерий, характеризующий степень достижения цели. Можно принять следующее решение: очень вкусный пирог получает отметку 5, просто вкусный пирог — отметку 4 и т. д.

Как вы полагаете, можно ли после такого решения переходить к оптимизации процесса?

Давайте разберемся. Нам важно количественно оценить результат оптимизации. Решает ли отметка эту задачу? Конечно, потому что, как мы договорились, отметка 5 соответствует очень вкусному пирогу и т. д. Другое дело, что этот подход, называемый ранговым, часто оказывается грубым, нечувствительным. Но возможность такой количественной оценки результатов не должна вызывать сомнений.

Другие примеры рангового подхода: определение чемпиона мира по фигурному катанию или гимнастике, дегустация вин, сравнение произведений искусства и т. д. Пример из области химии: сравнение продуктов по цвету, прозрачности, форме кристаллов.

Следующее требование: *параметр оптимизации должен выражаться одним числом*. Иногда это получается естественно, как регистрация показания прибора. Например, скорость движения машины определяется числом на спидометре. Чаше приходится производить некоторые вычисления. Так бывает при расчете выхода реакции. В химии часто требуется получить продукт с заданным отношением компонентов, например, $A : B = 3 : 2$. Один из возможных вариантов решения подобных задач состоит в том, чтобы выразить отношение одним числом (1,5) и в качестве параметра оптимизации пользоваться значениями отклонений (или квадратов отклонений) от этого числа.

Еще одно требование, связанное с количественной природой параметра оптимизации, — **однозначность** в статистическом смысле. Заданному набору значений факторов должно соответствовать одно, с точностью до ошибки эксперимента, значение параметра оптимизации. (Однако обратное неверно: одному и тому же значению параметра могут соответствовать разные наборы значений факторов.)

Для успешного достижения цели исследования необходимо, чтобы параметр оптимизации действительно оценивал эффективность функционирования системы в заранее выбранном смысле. Это требование является главным, определяющим корректность постановки задачи. «Если мы требуем победы и не знаем, что подразумеваем под этим, мы встретимся с призраком, стучащимся к нам в дверь», говорил Н.Винер.

Представление об эффективности не остается постоянным в ходе исследования. Оно меняется по мере накопления информации и в зависимости от достигнутых результатов. Это приводит к последовательному подходу при выборе параметра оптимизации. Так, например, на первых стадиях исследования технологических процессов в качестве параметра оптимизации часто используется выход продукта. Однако в дальнейшем, когда возможность повышения выхода исчерпана, нас начинают интересовать такие параметры, как себестоимость, чистота продукта и т. д.

Говоря об оценке эффективности функционирования системы, важно помнить, что речь идет о системе в целом. Часто система состоит из ряда подсистем, каждая из которых может оцениваться своим локальным параметром оптимизации. При этом оптимальность каждой из подсистем по своему параметру оптимизации «не исключает возможности гибели системы в целом».

Пример 3. При флотации сульфидной руды в лабораторных условиях изучалась эффективность применения нового реагента-пенообразователя по схеме рис. 3.2.

В качестве параметра оптимизации выбрано извлечение (при заданном качестве) концентрата в основной флотации. После проведения эксперимента выяснилось, что реагент дает более высокий выход концентрата по сравнению с прежним пенообразователем. Как вы считаете, обоснованно ли выбран параметр оптимизации, если ставилась задача оптимизации всего процесса флотационного обогащения руды?

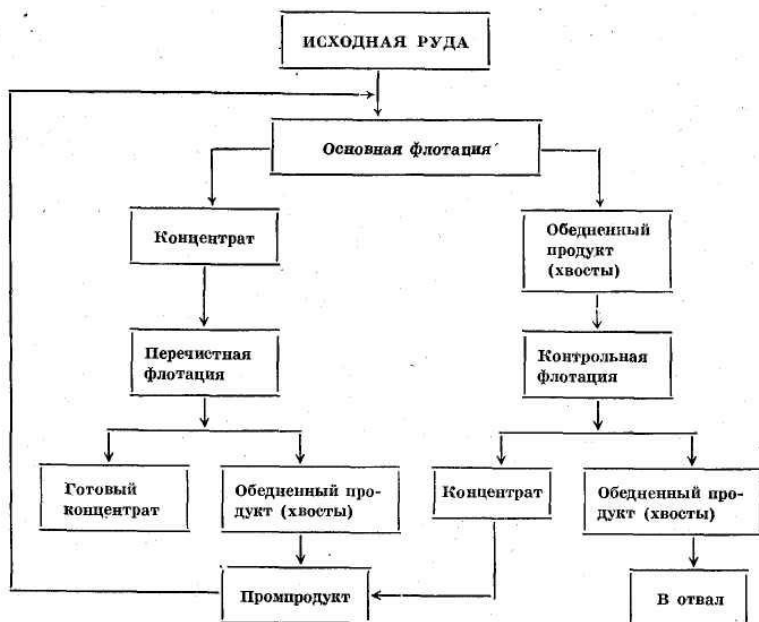


Рис. 3.2. Схема флотационного обогащения руды

Параметр оптимизации — извлечение концентрата в основной флотации — при решении задачи оптимизации всего процесса флотационного обогащения руды выбран не совсем обоснованно. Это правильный ответ, потому что существенно достижение конечной цели — получение готового концентрата (после перечистой флотации), а выбранный параметр оптимизации характеризует эффективность достижения промежуточной цели. Промежуточная цель — повышение выхода концентрата после основной флотации — была достигнута, но при промышленных испытаниях снизились показатели контрольной флотации, что привело к снижению извлечения и качества концентрата по всему циклу. Параметр оптимизации оказался неэффективным с точки зрения достижения конечной цели.

На эту сторону параметра оптимизации обращает внимание Вира: «Отличительной особенностью любой кибернетической системы можно считать полную бессмысленность рассмотрения ее иначе, как единого организма».

Мало иметь эффективный параметр оптимизации. Надо еще, чтобы он был эффективным в статистическом смысле. Понятие ста-

статистической эффективности достаточно сложное. Фактически это требование сводится к выбору параметра оптимизации, который определяется с наибольшей возможной точностью. (Если и эта точность недостаточна, тогда приходится обращаться к увеличению числа повторных опытов.)

Пусть, например, нас интересует исследование прочностных характеристик некоторого сплава. В качестве меры прочности можно использовать как прочность на разрыв, так и макротвердость. Поскольку эти характеристики функционально связаны, то с точки зрения эффективности они эквивалентны. Однако точность измерения первой характеристики существенно выше, чем второй. Требование статистической эффективности заставляет отдать предпочтение прочности на разрыв.

Следующее требование к параметру оптимизации — требование **универсальности** или **полноты**. Под универсальностью параметра оптимизации понимается его способность всесторонне характеризовать объект. В частности, технологические параметры оптимизации недостаточно универсальны: они не учитывают экономику. Универсальностью обладают, например, обобщенные параметры оптимизации, которые строятся как функции от нескольких частных параметров .

Пример выбора параметра оптимизации, обладающего полнотой, рассмотрен ниже для процессов зонной перекристаллизации. Обычно применяемый для этой цели коэффициент распределения, представляющий отношение концентраций примесей в твердой и жидкой фазах, излишне специфичен. Предложен более полный параметр оптимизации — **энтропийная функция S**

$$S = - \sum_{i=1}^m \sum_{j=1}^n c_{ij} \log c_{ij},$$

где c_{ij} — концентрация i -й примеси (при их числе m) в j -м участке слитка (при их числе n).

Желательно, чтобы параметр оптимизации имел *физический смысл, был простым и легко вычисляемым*.

Требование физического смысла связано с последующей интерпретацией результатов эксперимента. Не представляет труда объяснить, что значит максимум извлечения, максимум содержания ценного компонента. Эти и подобные им технологические параметры оптимизации имеют ясный физический смысл, но иногда для них может не выполняться, например, требование статистической эффективности. Тогда рекомендуется переходить к преобразованию

параметра оптимизации. Преобразование, например типа $\arcsin\sqrt{y}$, может сделать параметр оптимизации статистически эффективным (например, дисперсии становятся однородными), но остается неясным: что же значит достигнуть экстремума этой величины?

Второе требование часто также оказывается весьма существенным. Для процессов разделения термодинамические параметры оптимизации более универсальны. Однако на практике ими пользуются мало: их расчет довольно труден.

Из этих двух требований первое является более существенным, потому что часто удается найти идеальную характеристику системы и сравнить ее с реальной характеристикой. Иногда при этом целесообразно нормировать параметр с тем, чтобы он принимал значения от нуля до единицы.

Кроме высказанных требований и пожеланий при выборе параметра оптимизации нужно еще иметь в виду, что параметр оптимизации в некоторой степени оказывает влияние на вид математической модели исследуемого объекта. Экономические параметры, в силу их аддитивной природы, легче представляются простыми функциями, чем физико-химические показатели. Не случайно методы линейного программирования, основанные на простых моделях, получили широкое распространение именно в экономике. Температура плавления сплава является, как известно, сложной, многоэкстремальной характеристикой состава, тогда как стоимость сплава зависит от состава линейно. Итак, как видно, что найти параметр оптимизации, удовлетворяющий всем требованиям, задача не из легких. Существует два вида оптимизации объекта: структурная и параметрическая. При **структурной оптимизации** для достижения экстремальных значений целевой функции аппарат оптимизации применяют к самой структуре объекта. Так, например, если объектом исследования является технологический процесс (ТП), то **при структурной оптимизации осуществляют изменение последовательности технологических операций, методы проведения отдельных операций, состав и тип технологического оборудования** и т. п. В какой последовательности проводить обезгаживание деталей собранного мощного электровакуумного прибора (ЭВП), применять водородный или вакуумный отжиг, вести диффузию из жидкого или твердого источника, а может быть использовать ионную имплантацию вместо диффузии - **такие вопросы приходится решать при структурной оптимизации ТП.**

Очевидно, что **задачи структурной оптимизации возникают только на этапе проектировании или модернизации объекта.**

Чаще оптимизацию проводят при заданной структуре объекта, когда для достижения экстремума целевой функции можно менять только режимы (параметры) объекта. Такую оптимизацию называют *параметрической*.

При параметрической оптимизации объект представляют в виде «черного ящика» (рис. 3.3) с набором входных и выходных переменных, среди которых выделяют: y — вектор выходных переменных (параметры); x — вектор входных управляемых переменных (факторы), p — вектор входных контролируемых, но не управляемых переменных; ε — вектор случайных возмущений.

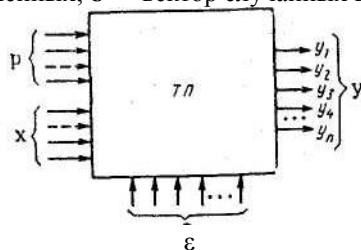


Рис. 3.3. Схема «черного ящика»

Тогда параметрическая оптимизация объекта заключается в нахождении и реализации таких способов варьирования управляемых переменных $x(t)$, при которых в рамках установленных ограничений на переменные y и x обеспечивается экстремум целевой функции, несмотря на изменения контролируемых переменных p и воздействие случайных возмущений ε .

При решении поставленной задачи оптимизации выделяют два этапа:

1) нахождение математического описания или модели объекта, т. е. такой совокупности математических величин y , x , p и их отношений, которая адекватно отображала бы свойства рассматриваемого объекта оптимизации;

2) разработка алгоритма оптимального управления, т.е. такой последовательности обработки текущей информации об объекте и такого порядка изменения уровней управляемых переменных x , которые обеспечивали бы поддержание целевой функции на экстремальном уровне при соблюдении ограничений и несмотря на воздействие неуправляемых факторов p и случайных возмущений ε .

Следует отметить, что между этими двумя этапами существует глубокая внутренняя взаимосвязь: с одной стороны, оптимальное управление невозможно без знания модели объекта, а с другой — выбор того или иного вида модели объекта во многом определяется тем, какой метод управления предполагается применить. Например,

использование алгоритма оптимизации Бокса— Вильсона (см. ниже) возможно только, если располагают локальным описанием целевой функции $y(x)$ вблизи текущей точки x . Однако если предполагается проводить оптимизацию методом Бокса-Вильсона, то нет необходимости строить сложную модель (например, полиномиальную высокого порядка) для описания $y(x)$ во всей области ограничений. Достаточно линейной модели, удовлетворительно описывающей $y(x)$ в малой окрестности текущей точки x .

3.1.3. Задачи с несколькими выходными параметрами

Задачи с одним выходным параметром имеют очевидные преимущества. Но на практике чаще всего приходится учитывать несколько выходных параметров. Иногда их число довольно велико. Так, например, при производстве резиновых и пластмассовых изделий приходится учитывать физико-механические, технологические, экономические, художественно-эстетические и другие параметры (прочность, эластичность, относительное удлинение, способность смеси прилипнуть к форме и т. д.). Математические модели можно построить для каждого из параметров, но *одновременно оптимизировать несколько функций невозможно*.

Обычно оптимизируется одна функция, наиболее важная с точки зрения цели исследования, при ограничениях, налагаемых другими функциями. Поэтому *из многих выходных параметров выбирается один в качестве параметра оптимизации, а остальные служат ограничениями*. Всегда полезно исследовать возможность уменьшения числа выходных параметров. Для этого можно воспользоваться *корреляционным анализом*.

При этом между всевозможными парами параметров необходимо вычислить *коэффициент парной корреляции*, который является общепринятой в математической статистике характеристикой связи между двумя случайными величинами. Если обозначить один параметр через y_1 , а другой — через y_2 , и число опытов, в которых они будут измеряться, — через N , так, что $u=1, 2, \dots, N$, где u — текущий номер опыта, то коэффициент парной корреляции r вычисляется по формуле

$$r_{y_1 y_2} = \frac{\sum_{u=1}^N (y_{1u} - \bar{y}_1) (y_{2u} - \bar{y}_2)}{\sqrt{\sum_{u=1}^N (y_{1u} - \bar{y}_1)^2 \sum_{u=1}^N (y_{2u} - \bar{y}_2)^2}}$$

Здесь

$$\bar{y}_1 = \sum_{u=1}^N y_{1u}/N \text{ и } \bar{y}_2 = \sum_{u=1}^N y_{2u}/N$$

средние арифметические соответственно для y_1 и y_2 .

Значения коэффициента парной корреляции могут лежать в пределах от -1 до $+1$. Если с ростом значения одного параметра возрастает значение другого, у коэффициента будет знак плюс, а если уменьшается, то минус. Чем ближе найденное значение r_{y_1, y_2} к единице, тем сильнее значение одного параметра зависит от того, какое значение принимает другой, т. е. между такими параметрами существует линейная связь, и при изучении процесса можно рассматривать только один из них. Необходимо помнить, что коэффициент парной корреляции как мера тесноты связи имеет четкий математический смысл только при *линейной зависимости* между параметрами и в случае *нормального их распределения*.

Для проверки значимости коэффициента парной корреляции нужно сравнить его значение с табличным (критическим) значением r , которое приведено в табл. 3.1.

Таблица 3. 1

Критические значения коэффициента парной корреляции при $\alpha=0,05$

Число степеней свободы f	Критическое значение r	Число степеней свободы f	Критическое значение r	Число степеней свободы f	Критическое значение r
1	0,997	9	0,602	17	0,456
2	0,950	10	0,576	18	0,444
3	0,878	11	0,553	19	0,433
4	0,811	12	0,532	20	0,423
5	0,754	13	0,514	30	0,349
6	0,707	14	0,497	50	0,273
7	0,666	15	0,482	80	0,217
8	0,632	16	0,468	100	0,195

Для пользования этой таблицей нужно знать число степеней свободы $f=N - 2$ и выбрать определенный уровень значимости, например, равный $0,05$. Такое значение уровня значимости называют еще 5%-ным уровнем риска, что соответствует вероятности верного ответа при проверке нашей гипотезы $P=1 - \alpha=0,95$, или 95%. Это значит, что в среднем только в 5% случаев возможна ошибка при проверке гипотезы.

В практических исследованиях 5%-ный уровень риска применяется наиболее часто. Но экспериментатор всегда свободен в выборе уровня значимости, и возможны ситуации, в которых, например, 1%-ный уровень риска. При этом возрастает надежность ответа. Проверка гипотезы сводится к сравнению абсолютной величины коэффициента парной корреляции с критическим значением. Если экспериментально найденное значение r меньше критического, то нет оснований считать, что имеется тесная линейная связь между параметрами, а если больше или равно, то гипотеза о корреляционной линейной связи не отвергается.

При высокой значимости коэффициента корреляции любой из двух анализируемых параметров можно исключить из рассмотрения как не содержащий дополнительной информации об объекте исследования. Исключить можно тот параметр, который технически труднее измерять, или тот, физический смысл которого менее ясен. При планировании эксперимента целесообразно измерять все параметры, затем оценить корреляцию между ними и строить модели для их минимально возможного числа или же воспользоваться обобщенным параметром. Но бывают случаи, когда приходится рассматривать и коррелированные параметры.

Подведем некоторые итоги.

Мы познакомились с некоторыми практически важными аспектами весьма сложной проблемы — выбора параметра оптимизации. Параметр оптимизации — это реакция (отклик) на воздействия факторов, которые определяют поведение изучаемой системы.

Параметры оптимизации бывают экономическими, технико-экономическими, технико-технологическими, статистическими, психологическими и т. д.

Параметр оптимизации должен быть:

- эффективным с точки зрения достижения цели;
- универсальным;
- количественным и выражаться одним числом;
- статистически эффективным;
- имеющим физический смысл, простым и легко вычисляемым;
- существующим для всех различных состояний.

В тех случаях, когда возникают трудности с количественной оценкой параметров оптимизации, приходится обращаться к ранговому подходу. В ходе исследования могут меняться априорные представления об объекте исследования, что приводит к последовательному подходу при выборе параметра оптимизации.

Из многих параметров, характеризующих объект исследования, только один, часто обобщенный, может служить параметром оптимизации. Остальные рассматриваются как ограничения.

3.2. Обобщенный параметр оптимизации

Единый параметр оптимизации можно получить путем обобщения нескольких (частных) параметров оптимизации. Как уже говорилось, из многих откликов, определяющих объект, очень часто трудно выбрать один, самый важный. Если же это возможно, тогда мы попадаем в ситуацию, описанную выше. Здесь же нам предстоит познакомиться с более сложной ситуацией, когда необходимо множество откликов обобщать (свертывать) в единый количественный признак. С таким обобщением связан ряд трудностей.

Каждый отклик имеет свой физический смысл и свою размерность. Чтобы объединить различные отклики, прежде всего приходится ввести для каждого из них некоторую *безразмерную шкалу*. Шкала должна быть однотипной для всех объединяемых откликов — это делает их сравнимыми. Выбор шкалы — не простая задача, зависящая от априорных сведений об откликах, а также от той точности, с которой мы хотим определить обобщенный признак.

После того как для каждого отклика построена безразмерная шкала, возникает следующая трудность — *выбор правила комбинирования исходных частных откликов в обобщенный показатель*. Единого правила здесь не существует. Можно идти различными путями, и выбор пути неформализован. Рассмотрим несколько различных способов построения обобщенного показателя.

3.2.1. Способы построения обобщенного отклика

Пусть исследуемый объект характеризуют n частных откликов y_u ($u=1,2, \dots, n$) и каждый из этих откликов измеряется в N опытах. Тогда y_{ui} — это значение u -го отклика в i -м опыте ($i=1,2, \dots, N$). Каждый из откликов y_u имеет свой физический смысл и, чаще всего, разную размерность. Введем простейшее преобразование: набор данных для каждого y_u поставим в соответствие с самым простым стандартным аналогом — шкалой, на которой имеется только два значения: 0 — брак, неудовлетворительное качество, 1 — годный продукт, удовлетворительное качество. Преобразованные значения обозначим так: y_{ui} — преобразованное значение u -го отклика в i -м опыте. Здесь мы применили шкалу, в которой использовано числовое множество из двух элементов (в данном случае 0 и 1). Стандартизовав таким образом шкалу частных откликов, мы подошли ко второму этапу — их

обобщению. По какому же правилу следует комбинировать частные отклики?

Будем рассуждать следующим образом. В ситуации, когда каждый преобразованный частный отклик принимает только два значения 0 и 1, желательнее, чтобы и обобщенный отклик принимал одно из этих двух возможных значений, причем так, чтобы значение 1 имело место, если, и только если, все частные отклики в этом опыте приняли значение 1. А если хотя бы один из откликов обратился в 0, то и обобщенный отклик будет нулем.

При таких рассуждениях для построения обобщенного отклика удобно воспользоваться формулой

$$Y_i = \sqrt[n]{\prod_{u=1}^n y_{ui}},$$

где Y_i — обобщенный отклик в i -м опыте; $\prod_{u=1}^n$ — произведение частных откликов $y_{1i} y_{2i} \dots y_{ni}$.

Корень мы ввели для того, чтобы связать эту формулу с другой, более сложной, которая будет рассмотрена далее. В данном же случае ничего не изменится, если написать

$$Y_i = \prod_{u=1}^n y_{ui}.$$

Совершенно очевидно, что такой подход слишком груб.

Рассмотрим пример.

Пример 1. При разработке оптимальной рецептуры нового пластифицированного полимерного материала качество продукции оценивалось семью выходными параметрами: y_1 — термостабильность композиции, мин; y_2 — блеск (баллы), y_3 — морозостойкость, °С; y_4 — модуль упругости при +20° С, кгс/см²; y_5 — предел прочности при растяжении, кгс/см²; y_6 — относительное удлинение при разрыве, %; y_7 — число перегибов до разрушения, шт.

Для каждого частного отклика введем следующие преобразования:

$$Y_{1i} = \begin{cases} 1, & \text{если } y_{1i} > 100; \\ 0, & \text{если } y_{1i} \leq 100; \end{cases} \quad Y_{2i} = \begin{cases} 1, & \text{если } y_{2i} < 20; \\ 0, & \text{если } y_{2i} \geq 20. \end{cases}$$

$$Y_{3i} = \begin{cases} 1, & \text{если } y_{3i} < -18; \\ 0, & \text{если } y_{3i} \geq -18; \end{cases} \quad Y_{4i} = \begin{cases} 1, & \text{если } y_{4i} < 120; \\ 0, & \text{если } y_{4i} \geq 120; \end{cases}$$

$$y_{5i} = \begin{cases} 1, & \text{если } y_{5i} > 200; \\ 0, & \text{если } y_{5i} \leq 200; \end{cases} \quad y_{6i} = \begin{cases} 1, & \text{если } y_{6i} > 200; \\ 0, & \text{если } y_{6i} \leq 200; \end{cases}$$

$$y_{7i} = \begin{cases} 1, & \text{если } y_{7i} > 25; \\ 0, & \text{если } y_{7i} \leq 25. \end{cases}$$

Данные для девяти опытов приведены в табл. 3.2.

Таблица 3.2

Натуральные, преобразованные и обобщенные отклики

Номер опыта	Натуральные частные отклики							Преобразованные частные отклики							Обобщенные отклики	
	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_1	Y_2
1	272	14	-25	103	215	299	103	1	1	1	1	1	1	1	1	1
2	187	20	-23	91	179	254	29	1	0	1	1	0	1	1	0	0
3	162	21	-24	102	216	270	99	1	0	1	1	1	1	1	0	1
4	461	14	-25	114	198	251	54	1	1	1	1	0	1	1	0	0
5	267	14	-21	105	208	268	31	1	1	1	1	1	1	1	1	1
6	250	24	-27	99	220	304	46	1	0	1	1	1	1	1	0	1
7	489	12	-25	123	201	238	33	1	1	1	0	1	1	1	0	1
8	380	14	-23	116	230	292	126	1	1	1	1	1	1	1	1	1
9	580	29	-22	100	215	304	48	1	0	1	1	1	1	1	0	1

Для комплексной оценки материалов были построены два обобщенных показателя:

- учитывающий требования и разработки и заказчика

$$Y_1 = \sqrt[7]{y_1 y_2 y_3 y_4 y_5 y_6 y_7},$$

- учитывающий только основные требования заказчика

$$Y_2 = \sqrt[3]{y_3 y_5 y_7},$$

По обобщенному показателю Y_1 могут рекомендоваться только три рецептуры (каждый опыт соответствует одной определенной рецептуре). Если же взять за основу только требования заказчика, то удовлетворительное качество имеют материалы в семи опытах.

Следует обратить внимание на еще один способ получения обобщенного отклика, который может применяться в тех случаях, когда для каждого из частных откликов известен «идеал», к которому нужно стремиться. Существует много способов введения метрики (**ввести метрику** — значит указать правило определения расстояний между любыми парами объектов из интересующего нас множества), задающей «близость к идеалу». Дополним предыдущие обозначения

еще одним: y_{u0} — наилучшее («идеальное») значение u -го отклика. Тогда $y_{ui} - y_{u0}$ можно рассматривать как некоторую меру близости к идеалу. Однако использовать разность при построении обобщенного отклика невозможно по двум причинам. Она имеет размерность соответствующего отклика, а у каждого из откликов может быть своя размерность, что препятствует, как мы уже знаем, их объединению. Отрицательный или положительный знак разности также создает неудобство. Чтобы перейти к безразмерным значениям, достаточно разность поделить на желаемое значение: $(y_{ui} - y_{u0})/y_{u0}$.

Чтобы **нивелировать знаки, можно разность возводить в квадрат**. Тогда обобщенный отклик получим по следующей формуле:

$$Y_i = \sum_{u=1}^n \left(\frac{y_{ui} - y_{u0}}{y_{u0}} \right)^2.$$

Если в некотором опыте все частные отклики совпадут с идеалом, то Y станет равным нулю. Это и есть то значение, к которому нужно стремиться. **Чем ближе к нулю, тем лучше**. Конечно, необходимо условиться о том, что считать нижней границей, если верхняя равна нулю. Следует обратить внимание на тот факт, что нуль здесь имеет другой смысл, чем в первом случае.

Среди недостатков такой оценки выделяется нивелировка частных откликов. Все они входят в обобщенный отклик на равных правах. На практике же различные показатели бывают далеко не равноправны. **Устранить этот недостаток можно введением некоторого веса a_u**

$$Y_i = \sum_{u=1}^n a_u \left(\frac{y_{ui} - y_{u0}}{y_{u0}} \right)^2,$$

причем

$$\sum_{u=1}^n a_u = 1 \text{ и } a_u > 0.$$

Чтобы проранжировать отклики по степени их важности и найти соответствующие веса, можно воспользоваться экспертными оценками.

Мы рассмотрели простейшие способы построения обобщенного показателя. Для перехода к более сложным способам нужно научиться фиксировать более тонкие различия на шкале преобразования откликов. Здесь в основном приходится опираться на опыт экспериментатора. Но, чтобы этот опыт разумно употребить в рамках формальных процедур, его тоже нужно формализовать. Наиболее

естественный путь такой формализации — *введение системы предпочтений экспериментатора на множестве значений каждого частного отклика, получение стандартной шкалы и затем обобщение результатов.*

Пользуясь системой предпочтений, можно получить более содержательную шкалу вместо шкалы классификаций с двумя классами. Пример построения такой шкалы мы рассмотрим ниже.

3.2.2. Шкала желательности (предпочтительности)

Одним из наиболее распространенных способов построения обобщенного отклика является обобщенная функция желательности Харрингтона. В основе построения этой обобщенной функции лежит идея преобразования натуральных значений частных откликов в безразмерную шкалу желательности или предпочтительности. Шкала желательности относится к психофизическим шкалам. Ее назначение — установление соответствия между физическими и психологическими параметрами. Здесь под **физическими параметрами** понимаются всевозможные отклики, характеризующие функционирование исследуемого объекта. Среди них могут быть эстетические и даже статистические параметры (см. классификацию параметров, приведенную выше), а под **психологическими параметрами** понимаются чисто субъективные оценки экспериментатора желательности (предпочтительности) того или иного значения отклика. Чтобы получить шкалу желательности, удобно пользоваться готовыми разработанными таблицами соответствий между отношениями предпочтения в эмпирической и числовой (психологической) системах (табл. 3.3).

Таблица 3.3

Стандартные отметки на шкале желательности

Желательность	Отметки на шкале желательности
Очень хорошо	1,00—0,80
Хорошо	0,80—0,63
Удовлетворительно	0,63—0,37
Плохо	0,37—0,20
Очень плохо	0,20—0,00

Значение частного отклика, переведенное в безразмерную шкалу желательности, обозначается через d_u ($u=1, 2, \dots, n$) и называется

частной желательностью (от desirable фр. — желательный). Шкала желательности имеет интервал от нуля до единицы. Значение $d_u=0$ соответствует абсолютно неприемлемому уровню данного свойства, а значение $d_u=1$ — самому лучшему значению свойства. Понятию «очень хорошо» соответствуют значения на шкале желательности $1 > d_u > 0,8$, а понятию «очень плохо» — $0 < d_u < 0,2$ и т. д. Выбор отметок на шкале желательности 0,63 и 0,37 объясняется удобством вычислений: $0,63 \approx 1 - (1/e)$, $0,37 \approx 1/e$. Значение $d_u = 0,37$ обычно соответствует границе допустимых значений.

В табл. 3.3 представлены числа, соответствующие некоторым точкам кривой (рис. 3.5), которая задается уравнением $d = e^{-e^{-y}}$ или $d = \exp[-\exp(-y)]$, где \exp — принятое обозначение экспоненты.

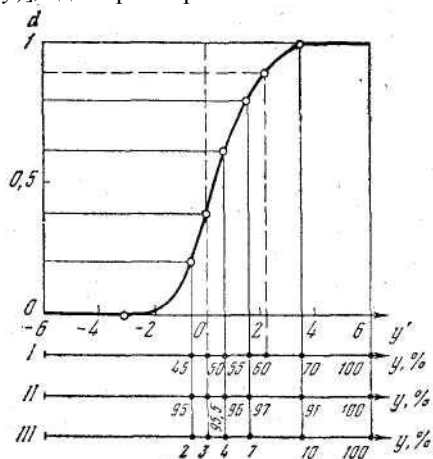


Рис. 3.5. Функция желательности

На оси ординат нанесены значения желательности, изменяющиеся от 0 до 1. По оси абсцисс указаны значения отклика, записанные в условном масштабе. За начало отсчета 0 по этой оси выбрано значение, соответствующее желательности 0,37. Выбор именно данной точки связан с тем, что она является точкой перегиба кривой, что в свою очередь создает определенные удобства при вычислениях. То же самое верно для значения желательности, соответствующего 0,63. Выбор этой кривой не является единственной возможностью. Однако она возникла в результате наблюдений за реальными решениями экспериментаторов и обладает такими полезными свойствами как **непрерывность, монотонность и гладкость**. Кроме того, эта кривая

хорошо передает тот факт, что в областях желательностей, близких к 0 и 1, «чувствительность» ее существенно ниже, чем в средней зоне.

Симметрично относительно нуля на оси y' (y' — кодированная шкала) расположены кодированные значения отклика. Значение на кодированной шкале принято выбирать от 3 до 6. Например, на рис. 3.5 использовано шесть интервалов в сторону убывания и шесть — в сторону возрастания. Ниже будет описан случай, когда выбрано три интервала. Выбор числа интервалов определяет крутизну кривой в средней зоне.

Пример 1. Пусть среди откликов будет выход реакции, естественные границы которого заключены между 0% и 100%.

Предположим, что 100% соответствует на шкале желательности единице, а 0% — нулю. Это отображено на рис. 3.5 на шкале, расположенной под шкалой y' . На ней пока имеются только две точки: 0 и 100. Выбор других точек зависит от ряда обстоятельств, таких, как сложившаяся в начальный момент ситуация, требования к результату, возможности экспериментатора.

Можно себе представить случай, когда процесс идет с выходом около 50%. О выходе более 70% трудно мечтать, а может быть он и невозможен из-за побочных реакций. Из сказанного ясно, что в данном случае выход в 70% для экспериментатора практически эквивалентен выходу в 100%. Это и отражено при выборе второй точки справа, для которой желательность близка к единице. Третья точка ограничивает область, которой экспериментатор может дать оценку «очень хорошо». В табл. 3.3 это соответствует диапазону 1,00—0,80. Какое натуральное значение y поставить в соответствие этой точке, зависит от «рискованности» или «осторожности» экспериментатора. Построение этой шкалы напоминает игру. Именно из теории игр заимствована эта терминология. Риск риском, но нужно еще и уметь играть. Опытный химик может сказать, что «больше 60% из этой реакции вы не выжмете». Тогда бы мы поставили в соответствие третьей точке выход 60%. Конечно, такое рассуждение может не иметь смысла, если вследствие ошибок эксперимента мы не в состоянии различить выхода в 60 и 70%. Следовательно, разным точкам на шкале должны соответствовать разные значения выхода. «Рискованный» экспериментатор может поставить в соответствие выход, например, в 67%. Область хороших результатов (0,80—0,63 по шкале желательности) экспериментатор заключил в границы 60—55%. Уже достигнутый результат — 50%, дает естественную нижнюю границу области «удовлетворительно», а результат ниже 45% не годится. На рис. 3.5 описанная ситуация обозначена (I).

Но может быть и другая ситуация. Имеется хорошо налаженный технологический процесс с выходом 95%, но возросли требования к чистоте продукта. Поэтому, наряду с другими мероприятиями, возникла задача увеличения выхода хотя бы до 98%. Эта ситуация на рисунке обозначена (II). Конечно, такая шкала возможна только при прецизионных измерениях. Совсем другая картина получается, когда речь идет о задаче синтеза нового вещества, которого до сих пор не удавалось получить в количествах, достаточных для идентификации. При выходе менее 2% нет способа в данном случае идентифицировать продукт. Любой выход выше 10% — превосходен. Эта ситуация (III), по-видимому, не нуждается в дальнейших комментариях.

Кривую желательности обычно используют как номограмму, поскольку это легко и оперативно. Так, если в ситуации (I) получен выход 63%, то ему будет соответствовать желательность 0,9, что показано пунктиром на рисунке. На практике такой простой прием часто оказывается вполне достаточным, но не всегда. Во-первых, потому что точность графического определения желательности может оказаться недостаточной, а, во-вторых, потому что эта точность зависит от положения на шкале y .

Тогда приходится прибегать к аналитическому методу определения желательности и с помощью шкалы y' и приведенного выше уравнения находить оценки желательности.

В нашем примере рассмотрен только один отклик — выход реакции, и границы его заданы естественно: от 0 до 100%. Однако, это не всегда бывает так. Если включить в рассмотрение отклики, характеризующие, например, качество материала, и границы станут весьма неопределенными.

Естественно возникает вопрос, на каком основании устанавливаются границы допустимых значений для частных откликов. При этом нужно иметь в виду, что ограничения могут быть односторонними в виде $y_u \leq y_{\max}$ или $y_u \geq y_{\min}$ и двусторонними в виде $y_{\min} \leq y_u \leq y_{\max}$. Здесь возможны две ситуации. Первая, самая благоприятная, возможна, если экспериментатор располагает инструкцией, в которой четко сформулированы требования ко всем частным откликам, т. е. имеется спецификация с одним или двумя ограничивающими пределами. Тогда отметка на шкале желательности $d_u=0,37$ соответствует y_{\min} , если имеется одностороннее ограничение $y_u \geq y_{\min}$ или y_{\max} для $y_u \leq y_{\max}$. В случае двустороннего ограничения этой отметке ставится в соответствие и y_{\min} , и y_{\max} . Во второй ситуации спецификация отсутствует, тогда ограничения на шкале и другие отметки делаются весьма субъективно, на основании опыта и интуиции экспериментатора.

Очевидно, что в таком случае довольно опасно руководствоваться мнением одного-единственного исследователя, а желательно учесть мнения нескольких специалистов. При обобщении ряда мнений и установлении степени согласованности между различными специалистами можно воспользоваться методом ранговой корреляции.

3.2.3. Преобразование частных откликов в частные функции желательности

Возьмем случай, когда существует спецификация с одним или двумя ограничивающими пределами и эти пределы являются единственными значениями качества. Тогда вне пределов $d_u=0$, а внутри пределов $d_u=1$.

Пусть y_{\min} — нижний предел спецификации и $y_u \geq y_{\min}$. При таком одностороннем ограничении частная функция желательности будет иметь вид

$$d_u = \begin{cases} 0, & \text{если } y_u < y_{\min}; \\ 1, & \text{если } y_u \geq y_{\min}. \end{cases}$$

Аналогично для двустороннего ограничения получаем

$$d_u = \begin{cases} 0, & \text{если } y_u < y_{\min} \text{ и } y_u > y_{\max}; \\ 1, & \text{если } y_{\min} \leq y_u \leq y_{\max}. \end{cases}$$

Таким образом, мы пришли к случаю, рассмотренному в 3.2.1: шкала желательности выродилась в простейшую шкалу классификаций с двумя классами эквивалентности. Эти два случая показаны на рис. 3.6. Но такое положение, когда ограничивающие пределы спецификации являются единственными критериями качества, встречается довольно редко, так как трудно разделить результаты твердой границей на две категории: годен, не годен. Поэтому преобразования частных откликов в их стандартные аналоги на шкале желательности осуществляются по более сложным законам. Примером такого более сложного преобразования служит таблица желательности (табл. 3.3) и соответствующая ей функция желательности.

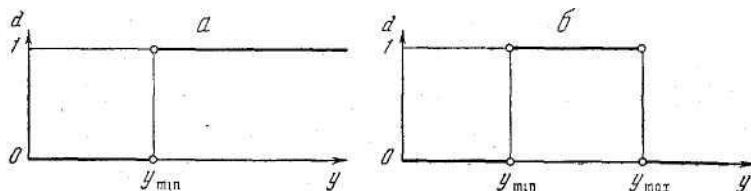


Рис. 3.6. Задание частной функции желательности

a — одностороннее ограничение; *б* — двустороннее ограничение

При односторонних ограничениях $y_i \leq y_{\max}$ или $y_i \geq y_{\min}$ на рис. 3.7 представлена частная функция желательности для свойства, ограниченного с одной стороны.

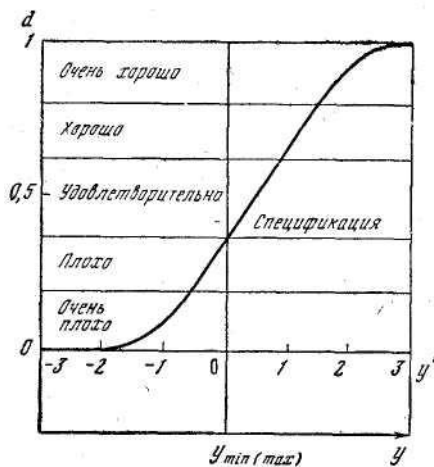


Рис. 3.7. Функция желательности для свойства, ограниченного с одной стороны

К числу свойств, подчиняющихся одностороннему ограничению, относятся очень многие характеристики качества материалов: теплостойкость, прочность, ударная вязкость, морозостойкость, модуль упругости, относительное удлинение при разрыве и т. д. (это можно было видеть из примера 1). Для всех этих показателей ограничение имело вид $y_i \geq y_{\min}$.

Другой вид одностороннего ограничения $y_i \leq y_{\max}$ характерен для таких показателей, как содержание различных вредных примесей, влажность, удельный вес, содержание дорогих или дефицитных компонентов и т. п.

Для двустороннего ограничения $y_{\min} \leq y_i \leq y_{\max}$ пример функции желательности показан на рис. 3.8.

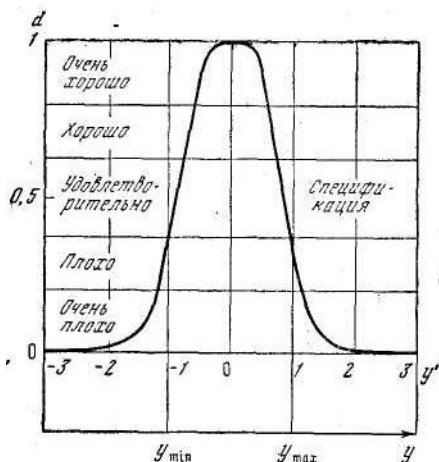


Рис. 3.8. Функция желательности для свойства, ограниченного с двух сторон

Случай двустороннего ограничения встречается не так часто, как одностороннего; и он более сложен для оценки откликов. Примерами могут служить молекулярный вес материалов, предназначенных для переработки литьем или экструзией, насыпной вес и вес утряски, индекс расплава и т. п. Индекс расплава, например, является важной характеристикой литьевых и экструзионных типов полиметилметакрилата. Слишком низкий индекс наблюдается у материалов с недостаточной текучестью, что отрицательно влияет на ход технологического процесса при переработке полимера. При высоком показателе индекса расплава возникают затруднения во время формирования материала. В некоторых сложных случаях к построению шкалы желательности приходится подходить итеративно.

Шкала желательности есть попытка формализации представлений экспериментатора о важности тех или других значений частных откликов. Нет никакой гарантии, что такие представления можно считать правильными. Особенно, если не существует спецификации. В тех случаях, когда шкала желательности создается совместно экспериментатором и консультантом по планированию эксперимента, можно представить себе и такую ситуацию, когда не было между ними достаточно глубокого взаимопонимания и шкала получилась неудачной. Тогда нужно учесть все недостатки и построить другую шкалу, более точно отражающую предпочтительность значений откликов.

3.2.4. Обобщенная функция желательности

После того, как выбрана шкала желательности и частные отклики преобразованы в частные функции желательности, можно приступить к основной задаче — построению обобщенного показателя D , названного Харрингтоном *обобщенной функцией желательности*. Обобщать, то есть переходить от d_i к D , предлагается по формуле

$$D = \sqrt[n]{\prod_{u=1}^n d_u}.$$

Здесь обобщенная функция желательности задается как среднее геометрическое частных желательностей. Такое представление можно рассматривать как удобную модель психологической реакции исследователя, которая возникает при решении определенного класса задач. Примером может служить установление пригодности материала с данным набором свойств для использования его в заданных условиях. Если хотя бы один частный отклик, входящий в комплекс параметров качества материала, не удовлетворяет требованиям спецификации (например, при определенной температуре материал становится хрупким и разрушается), то как бы ни были хороши прочие свойства, этот материал не может быть использован по назначению. Действительно, способ задания обобщенной функции желательности таков, что если хотя бы одна частная желательность $d_u=0$, то обобщенная функция тоже будет равна нулю, с другой стороны $D=1$ тогда и только тогда, когда все $d_u=1$ ($u=1, 2, \dots, n$). Обобщенная функция желательности весьма чувствительна к малым значениям частных желательностей. Можно представить себе другие задачи, когда столь жесткие свойства обобщенного критерия окажутся неприемлемыми. Тогда нужно перейти к другим способам обобщения. Вернемся к функции желательности.

Способ задания базовых отметок шкалы желательности, представленный в табл. 3.3, один и тот же как для частных желательностей, так и для обобщенной. Так, если $d_1, d_2, \dots, d_n=0,63$, то и $D=0,63$, если $d_1, d_2, \dots, d_n=0,37$, то и $D=0,37$ и т. п. В обобщенную функцию желательности могут входить самые разнообразные частные отклики: технологические, технико-экономические, экономические, эстетические и т. п.

В 3.2.1 мы рассмотрели в примере 1 построение обобщенного отклика с использованием шкалы классификации с двумя классами эквивалентности 0 и 1. Теперь для этих же исходных данных построим обобщенную функцию желательности.

Пример 2. Пусть полимерный материал характеризуют те же семь частных откликов, как и в примере 1. Используем более тонкий способ преобразования откликов в безразмерные величины, чем мы это сделали в предыдущем примере, и построим для них шкалу желательности (рис. 3.9).

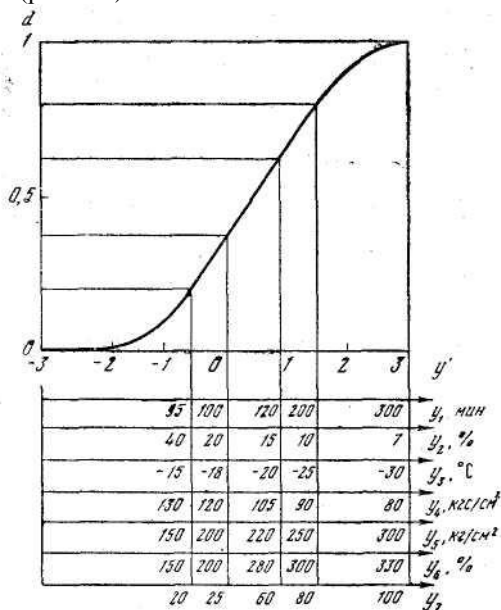


Рис. 3.9. Шкала желательности для нового полимерного материала

Натуральные значения частных откликов преобразуем в частные функции желательности (см. табл. 3.4).

По обобщенному критерию D_1 четыре рецептуры получили оценку хорошо, пять — удовлетворительно, а по критерию D_2 , учитывающего только требование заказчика, три рецептуры оценены хорошими, шесть удовлетворительными. Если сравнить эти данные с примером 1, в котором шкала принимала только два значения (0 — брак и 1 — годный продукт), то становятся совершенно очевидными преимущества функции желательности как более тонкого обобщенного критерия.

Мы рассмотрели самый простой способ получения обобщенного критерия — *графический*. Но часто приходится прибегать к аналитической зависимости, особенно при систематических исследованиях одного и того же объекта. Тогда можно использовать

полиномы невысоких степеней, обычно первой или второй степени. Проверка пригодности того или иного полинома является специальной статистической задачей. Решение этой задачи будет рассмотрено несколько позже.

Располагая информацией о y'_i мы можем рассчитать d_i по уже упоминавшейся формуле

$$d_i = \exp[-\exp(-y'_i)].$$

Для получения кодированных значений y'_i в данном примере взяты три равномерных интервала, т. е. выбран код: $-3, -2, -1, 0, +1, +2, +3$. (Кодированные значения откликов относятся к верхней границе интервала).

Такой код можно поставить в соответствие абсциссе однажды построенной кривой. Если требуется регулировать ее крутизну, то это можно сделать изменением числа интервалов. Чтобы приспособить такую стандартную кривую к реальным частным откликам, достаточно дополнить табл. 3.3 следующей информацией:

d_i	y'_i	y_1	y_2	y_3	y_4	y_5	y_6	y_7
1,00—0,80	3,0	300	7	-30	80	300	330	100
0,80—0,63	1,5	200	10	-25	90	250	280	80
0,63—0,37	0,85	120	15	-20	105	220	250	60
0,37—0,20	0,00	100	20	-18	120	200	200	25
0,20—0,00	-0,50	95	40	-15	130	150	150	20

Обобщенная функция желательности является некоторым абстрактным построением и поэтому, прежде чем рекомендовать ее в качестве единого критерия оптимизации, представлялось интересным исследовать такие ее важные свойства, как *адекватность, статистическая чувствительность и эффективность*. Оказалось, что для шкал желательности статистическая чувствительность и эффективность частных и обобщенной функций желательности не ниже, чем таковые для любого технологического показателя, им соответствующего.

Обобщенная функция желательности является количественным, однозначным, единым и универсальным показателем качества исследуемого объекта, и если добавить еще такие свойства, как адекватность, эффективность и статистическая чувствительность, то становится ясным, что ее можно использовать в качестве критерия оптимизации.

Обобщенная функция желательности нашла широкое применение для оценки качества полимерных материалов, резиновых и латексных изделий, а также при разработке различных рецептов. Широко используется она также и в промышленности.

Таблица 3.4

Натуральные и обобщенные по функции желательности отклики

Номер опыта	Натуральные значения откликов							Частные желательности							D ₁	D ₂	Оценка по шкале желательности	
	y ₁	y ₂	y ₃	y ₄	y ₅	y ₆	y ₇	d ₁	d ₂	d ₃	d ₄	d ₅	d ₆	d ₇				
	1	272	14	-25	103	215	299	103	0.98	0.67	0.80	0.71	0.55	0.80				1.00
2	187	20	-23	91	179	254	29	0.77	0.36	0.73	0.79	0.32	0.53	0.29	0.500	ул.	0.407	ул.
3	162	21	-24	102	216	270	99	0.68	0.35	0.75	0.72	0.55	0.56	0.99	0.630	хор.	0.742	хор.
4	461	14	-25	114	198	251	54	1.00	0.67	0.80	0.47	0.38	0.53	0.57	0.603	ул.	0.558	ул.
5	257	14	-21	105	208	268	31	0.97	0.67	0.70	0.63	0.44	0.54	0.39	0.595	ул.	0.493	ул.
6	250	24	-27	99	220	304	46	0.95	0.32	0.90	0.70	0.63	0.81	0.48	0.646	хор.	0.621	ул.
7	489	12	-25	123	201	238	33	1.00	0.72	0.80	0.35	0.38	0.50	0.40	0.533	ул.	0.495	ул.
8	380	14	-23	116	230	292	126	1.00	0.67	0.73	0.44	0.67	0.79	1.00	0.733	хор.	0.788	хор.
9	580	29	-22	100	215	304	48	1.00	0.30	0.72	0.70	0.55	0.81	0.52	0.626	ул.	0.603	ул.

$$D_1 = \sqrt[7]{d_1 \cdot d_2 \cdot d_3 \cdot d_4 \cdot d_5 \cdot d_6 \cdot d_7}, \quad D_2 = \sqrt[3]{d_5 \cdot d_6 \cdot d_7}$$

В заключение подведем некоторые итоги.

Построение обобщенного параметра оптимизации связано с созданием единого признака, количественно определяющего функционирование исследуемого объекта с многими выходными параметрами. При этом возникают некоторые трудности. Каждый *выходной параметр*—*отклик*—имеет свой физический смысл, свою размерность. Чтобы объединить различные отклики, необходимо ввести единую для всех откликов искусственную метрику. Набор данных каждого отклика нужно поставить в соответствие с некоторым стандартным аналогом, с безразмерной шкалой. Поэтому первым вопросом, который нужно решить при построении обобщенного параметра оптимизации, является вопрос о выборе шкалы. ***Шкала должна быть однотипной для всех объединяемых откликов.*** Построение шкалы во многом зависит от уровня априорных сведений о выходных параметрах, а также от той точности, с которой мы хотим определить обобщенный отклик.

Второй важный вопрос — выбор правила комбинирования исходных частных откликов в обобщенный показатель. Единого правила не существует, здесь можно идти различными путями, и выбор пути неформализован. Рассмотрено несколько различных способов построения обобщенного показателя.

3.3. Факторы

Теперь рассмотрим способы воздействия на оптимизируемый объект. Ранее мы определили, что ***способы воздействия на объект были названы факторами.***

После того как выбран объект исследования и параметр оптимизации, нужно включить в рассмотрение все существенные факторы, которые могут влиять на процесс. Если какой-либо существенный фактор окажется неучтенным, то это может привести к непредсказуемым последствиям. Так, если неучтенный фактор произвольно *флуктуировал* — *принимал случайные значения, которые экспериментатор не контролировал*, — это значительно увеличит ошибку опыта. При поддержании фактора на некотором фиксированном уровне может быть получено ложное представление об оптимуме, так как нет гарантии, что фиксированный уровень является оптимальным.

Ранее отмечалось, что число различных состояний объекта p^k , где p — число уровней, а k — число факторов. При этом возникает вопрос, а как же преодолеть большое число опытов? Чем больше факторов, тем больше опытов. Действительно, число опытов растет по показательной

функции. Размерность факторного пространства увеличивается, и математики в таких случаях говорят о «проклятии размерности».

Рекомендации о том, как преодолеть «проклятие размерности», будут рассмотрены позже.

Если число факторов больше пятнадцати, нужно обратиться к методам отсеивания несущественных факторов. Здесь можно воспользоваться формализацией априорной информации, методом случайного баланса, планами Плаккета-Бермана и др. Иногда эти планы применяются и при меньшем числе факторов.

3.3.1. Определение фактора

Фактором называется измеряемая переменная величина, принимающая в некоторый момент времени определенное значение. Факторы соответствуют способам воздействия на объект исследования. Так же, как и параметр оптимизации, каждый фактор имеет **область определения**. Мы будем считать фактор заданным, если вместе с его названием указана область его определения. **Под областью определения** понимается совокупность всех значений, которые в принципе может принимать данный фактор. Ясно, что совокупность значений фактора, которая используется в эксперименте, является подмножеством из множества значений, образующих область определения.

Область определения может быть **непрерывной и дискретной**. Так, для факторов с непрерывной областью определения, таких, как температура, время, количество вещества и т. п., всегда **выбираются дискретные множества уровней**. В практических задачах области определения факторов, как правило, ограничены. Ограничения могут носить принципиальный либо технический характер.

Произведем классификацию факторов в зависимости от того, является ли фактор переменной величиной, которую можно оценивать количественно: измерять, взвешивать, титровать и т. п., или же он — некоторая переменная, характеризующаяся качественными свойствами.

Факторы разделяются на количественные и качественные. **Качественные факторы** — это разные вещества, разные технологические способы, аппараты, исполнители и т. д.

Хотя качественным факторам не соответствует числовая шкала в том смысле, как это понимается для количественных факторов, однако можно построить условную порядковую шкалу, которая ставит в соответствие уровням качественного фактора числа натурального ряда, т. е. производит кодирование. Порядок уровней может быть произволен, но после кодирования он фиксируется.

В ряде случаев граница между понятием качественного и количественного фактора весьма условна. Пусть, например, при изучении воспроизводимости результатов химического анализа надо установить влияние положения тигля с навеской в муфельной печи. Можно разделить под печи на квадраты и считать номера квадратов уровнями качественного фактора, определяющего положение тигля. Вместо этого можно ввести два количественных фактора — ширину и длину пода печи. Качественным факторам не соответствует числовая шкала, и порядок уровней факторов не играет роли.

Время реакции, температура, концентрация реагирующих веществ, скорость подачи веществ, величина рН — это примеры наиболее часто встречающихся **количественных факторов**. Различные реагенты, адсорбенты, вулканизирующие агенты, кислоты, металлы являются примером уровней качественных факторов.

Пример 1. Этот пример относится к исследованию процесса вулканизации бутадиен-стирольного каучука солями непердельных кислот. В планирование эксперимента были включены следующие факторы: \tilde{x}_1 — температура вулканизации, °С; \tilde{x}_2 — время вулканизации, мин; \tilde{x}_3 — количество инициатора, вес. ч.; \tilde{x}_4 — количество вулканизирующего агента, вес. ч; \tilde{x}_5 — количество окисла, вес. ч.; \tilde{x}_6 — тип окисла (окись цинка или окись магния); \tilde{x}_7 — тип кислотного остатка (метакрилат, малеат); \tilde{x}_8 — тип катиона соли (Na, Mg).

Довольно большое количество факторов и наличие среди них качественных объясняется тем, что планировать эксперимент приходилось на первой стадии, когда еще неясно, какие вещества нужно использовать для структурирования эластомеров, какие соли добавить для увеличения прочности, какие ускорители окажутся наиболее эффективными и т. д. .

Пример 2. Многостадийный процесс получения ацетилацетона характерен тем, что приходилось изучать факторы, влияющие на четыре стадии процесса, так как у определялся только в конце 4-й стадии (рис. 3.10).

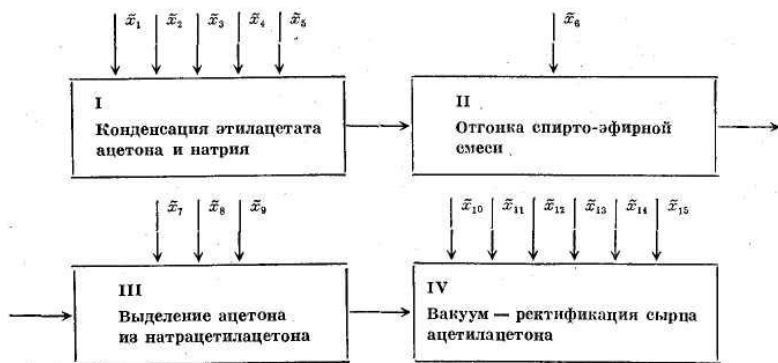


Рис. 3.10. Процесс получения ацетилацетона

I. Стадия конденсации. \tilde{x}_1 — температура реакции конденсации, °С; \tilde{x}_2 — время прилива ацетона, мин; \tilde{x}_3 — время выдержки, час; \tilde{x}_4 — соотношение компонентов, z/z ; \tilde{x}_5 — скорость перемешивания, об/сек.

II. Стадия отгонки спирто-эфирной смеси. \tilde{x}_6 — конечная температура сухого остатка, °С.

III. Стадия выделения ацетилацетона из натрацетилацетона. \tilde{x}_7 — величина pH; \tilde{x}_8 — скорость подачи соляной кислоты, мл/сек; \tilde{x}_9 — температура при выделении, °С.

IV. Стадия вакуум-ректификации сырца ацетилацетона. \tilde{x}_{10} — температура отгонки спирто-эфирной смеси 1-й фракции, °С; \tilde{x}_{11} — температура отгонки спирто-эфирной смеси 2-й фракции, °С; \tilde{x}_{12} — температура отгонки спирто-эфирной смеси 3-й фракции, °С; \tilde{x}_{13} — время отгонки 1-й фракции, мин; \tilde{x}_{14} — время отгонки 2-й фракции, мин; \tilde{x}_{15} — время отгонки 3-й фракции, мин.

Рассмотрим вопрос, можно ли включать в планирование эксперимента факторы, относящиеся к различным стадиям?

В рассматриваемом примере процесс получения ацетилацетона, состоящий из четырех стадий, удобно представить как единое целое в виде одного «черного ящика» (рис. 3.11). На этот «черный ящик» воздействуют 15 факторов. Почему возникла необходимость рассматривать все стадии как единое целое?

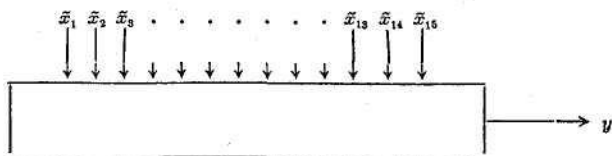


Рис. 3.11. Процесс получения ацетилацетона

Параметр оптимизации измеряется в конце последней стадии. На предыдущих стадиях выходной параметр не измеряется: отсутствуют нужные аналитические методики. Возможны и другие причины, например, параметр оптимизации отдельной стадии противоречит общей цели оптимизации. Но было бы неправильным считать, что во всех случаях при оптимизации многостадийных процессов нужно рассматривать все стадии как единое целое. Весьма часто оптимизация отдельных стадий вполне оправдана и очевидна.

Так, процесс получения сульфадимизина состоит из трех химических стадий: получения сульгина, получения ацетилацетона (эта стадия, как видно из рассматриваемого примера, в свою очередь состоит из четырех частей) и конденсации сульгина с ацетилацетоном. Сульгин и ацетилацетон имеют самостоятельное значение. Процессы их получения выделены в отдельные производства, зачастую территориально не объединенные. Оптимизировать совместно все три стадии не представляется возможным и целесообразным.

Таким образом, в планирование эксперимента можно включать факторы, относящиеся к различным стадиям, но не во всех случаях это является необходимым.

3.3.2. Требования, предъявляемые к факторам при планировании эксперимента

Выше дано определение понятию «фактор» и приведены примеры факторов. Теперь сформулируем требования, предъявляемые к факторам.

При планировании эксперимента факторы должны быть *управляемыми*. Это значит, что экспериментатор, выбрав нужное значение фактора, может его поддерживать постоянным в течение всего опыта, т. е. может управлять фактором. В этом состоит особенность «активного» эксперимента. **Планировать эксперимент можно только в том случае, если уровни факторов подчиняются воле экспериментатора.**

Представьте себе, что вы изучаете процесс синтеза аммиака. Колонна синтеза установлена на открытой площадке. Является ли температура воздуха фактором, который можно включить в планирование эксперимента?

Температура воздуха — фактор неуправляемый. Мы еще не научились делать погоду по заказу. А в планировании могут участвовать только те факторы, которыми можно управлять, — устанавливать и поддерживать на выбранном уровне в течение опыта или менять по заданной программе. Температурой окружающей среды в данном случае управлять невозможно. ***Ее можно только контролировать.***

Чтобы ***точно определить фактор***, нужно указать ***последовательность действий (операций)***, с помощью которых ***устанавливаются его конкретные значения (уровни)***. Такое определение фактора будем называть ***операциональным***. Так, если фактором является давление в некотором аппарате, то совершенно необходимо указать, в какой точке и с помощью какого прибора оно измеряется и как оно устанавливается. Введение операционального определения обеспечивает однозначное понимание фактора.

С операциональным определением связаны выбор ***размерности фактора и точность его фиксирования***. Мы привыкли считать, что выбор размерности фактора не представляет особой трудности. Экспериментатор хорошо ориентируется в том, какую размерность нужно использовать. Это действительно так в тех случаях, когда существует устоявшаяся традиция, построены измерительные шкалы, приборы, созданы эталоны и т. д. Так обстоит дело при измерении температуры, времени, давления и т. д. Но бывает, что выбор размерности превращается в весьма трудную проблему выбора измерительных шкал. ***Замена одной измерительной шкалы другой называется преобразованием шкал***. Оно может быть использовано для упрощения модели объекта.

Точность замера факторов должна быть возможно более высокой. Степень точности определяется диапазоном изменения факторов. При изучении процесса, который длится десятки часов, нет необходимости учитывать доли минуты, а в быстрых процессах необходимо учитывать, быть может, доли секунды. Если факторы измеряются с большой ошибкой или особенность объекта исследования такова, что значения факторов трудно поддерживать на выбранном уровне (уровень фактора «плывет»), то экспериментатору следует обратиться к ***конфлюэнтному анализу***.

Факторы должны быть непосредственными воздействиями на объект. Факторы должны быть *однозначны*. Трудно управлять фактором, который является функцией других факторов. Но в планировании могут участвовать *сложные факторы*, такие, как соотношения между компонентами, их логарифмы и т. п.

Необходимость введения сложных факторов возникает при желании представить *динамические особенности объекта в статической форме*. Пусть, например, требуется найти оптимальный режим подъема температуры в реакторе. Если относительно температуры известно, что она должна нарастать линейно, то в качестве фактора *вместо функции* (в данном случае линейной) можно использовать тангенс угла наклона, т. е. *градиент*. Положение усложняется, когда исходная температура не зафиксирована. Тогда ее приходится вводить в качестве еще одного фактора. Для более сложных кривых пришлось бы ввести большее число факторов (производные высоких порядков, координаты особых точек и т. д.). Поэтому целесообразно пользоваться сложным качественным фактором — номером кривой. *Различные варианты кривых рассматриваются в качестве уровней*. Это могут быть разные режимы термообработки сплавов, переходные процессы в системах управления и т. д. Мы показали, как можно сложный фактор-функцию представить с помощью простых однозначных факторов.

Пример 3. При оптимизации процесса получения одного производного пиперазина изучалось влияние семи факторов, среди которых были соотношения между компонентами: \tilde{x}_1 — количество едкого натра, г/мол; \tilde{x}_2 — способ поддержания рН; \tilde{x}_3 — время прилива вещества *a*, час; \tilde{x}_4 — время выдержки реакционной массы, час; \tilde{x}_5 — температура реакционной среды, °С; \tilde{x}_6 — весовое соотношение вещества *b* и метанола, г/г; \tilde{x}_7 — мольное соотношение вещества *a* и вещества *b*, г/мол/г/мол.

Изучается процесс растворения твердого тела в жидкости — диффузионный процесс. Необходимо ответить на вопрос, может ли скорость диффузии служить фактором в планировании эксперимента?

Скорость диффузии зависит от концентрации, величины поверхности соприкосновения двух фаз — жидкости и твердого тела, от коэффициента растворения. Коэффициент растворения зависит от коэффициента диффузии и от толщины диффузионного слоя. Коэффициент диффузии, в свою очередь, является функцией нескольких переменных.

Конкретное значение скорости диффузии определяется сочетанием значений других факторов. Если бы мы могли управлять скоростью диффузии, придавая ей в каждом опыте желаемое значение, то она могла бы стать фактором. Скорость диффузии не может являться фактором при планировании эксперимента, так как управлять ею практически невозможно.

При растворении твердого тела в жидкости образуется диффузионный слой, прилегающий к поверхности твердого тела. Состав этого слоя неодинаков в различных зонах. В пограничной части слой в большей или меньшей степени находится в состоянии равновесия, и концентрация растворенного вещества в нем приближается к концентрации насыщенного раствора, $c_{\text{нас}}$. В части слоя, прилегающей к внутреннему объему жидкости, концентрация растворенного вещества приближается к концентрации c в остальном объеме жидкости. Скорость диффузии тем больше, чем больше различие в концентрациях ($c_{\text{нас}} - c$) диффундирующего вещества

$$\frac{dc}{dt} = KS (c_{\text{нас}} - c),$$

где dc/dt — скорость изменения концентрации в объеме рассматриваемой фазы, S — величина поверхности соприкосновения данных фаз, K — коэффициент растворения.

Коэффициент растворения зависит от коэффициента диффузии D растворяемого вещества и от толщины диффузионного слоя $\delta K = D/\delta$. Коэффициент диффузии, в свою очередь, зависит от ряда факторов.

Теперь стало ясно, что скорость диффузии является функцией многих переменных. Скоростью диффузии весьма трудно управлять.

Такие требования, предъявляемые к факторам при планировании эксперимента, как *управляемость* и *однозначность*, в рассматриваемом примере не выполняются.

3.3.3. Требования к совокупности факторов

При планировании эксперимента обычно одновременно изменяется несколько факторов. Поэтому очень важно сформулировать требования, которые предъявляются к совокупности факторов. Прежде всего выдвигается требование *совместимости*. Совместимость факторов означает, что все их комбинации осуществимы и безопасны. Это очень важное требование. Если не уделять достаточного внимания на требование совместимости факторов то можно запланировать такие условия опыта, которые могут привести к взрыву установки или осмолению продукта. Такой результат очень далек от целей оптимизации.

Несовместимость факторов может наблюдаться на границах областей их определения. Избавиться от нее можно сокращением областей. Положение усложняется, если несовместимость проявляется внутри областей определения. Одно из возможных решений — разбиение на подобласти и решение двух отдельных задач.

При планировании эксперимента важна *независимость* факторов, т. е. возможность установления фактора на любом уровне вне зависимости от уровней других факторов. Если это условие невыполнимо, то невозможно планировать эксперимент. Итак, мы подошли ко второму требованию — *отсутствию корреляции между факторами*. Требование некоррелированности не означает, что между значениями факторов нет никакой связи. Достаточно, чтобы *связь не была линейной*.

Исследуется некоторая термодинамическая система. Можно ли включить в планирование эксперимента следующие три фактора: \tilde{x}_1 — давление, атм, \tilde{x}_2 — объем, л, \tilde{x}_3 — температуру, °К?

Пусть в термодинамической системе имеет место уравнение Менделеева—Клапейрона $PV=nRT$ и заданы два фактора, например $V(x_1)$ и $T(x_2)$. Тогда $P(x_3)$ может быть вычислено. То же самое и с двумя другими парами факторов. Поэтому в планирование можно включать два (а не три) фактора. Здесь возможны три комбинации: 1) x_1 и x_2 , 2) x_1 и x_3 , 3) x_2 и x_3 .

Ранее мы уже приводили примеры факторов, относящиеся к органической химии и полимерам. А теперь перейдем к другим областям, памятуя изречение Ньютона, что примеры так же поучительны, как и правила.

Области практических приложений планирования эксперимента чрезвычайно многообразны: химия, металлургия, биология, медицина, обогащение полезных ископаемых, пищевая и текстильная промышленность, сельское хозяйство, военное дело и др. .

Применяется планирование эксперимента и в несколько неожиданных областях исследования, в таких, как геронтология (наука о долголетьи), при классификации образцов древней керамики, в хлебопечении и табачном деле.

В зависимости от объектов исследования меняются и факторы. В приведенных примерах, не возможно отразить все это многообразие. Остановимся на более типичных для практики случаях.

В выборе примеров мы также руководствовались принципом многофакторности; приводили задачи, в которых количество факторов было бы не меньше четырех, так как придумать примеры с двумя-тремя факторами очень легко может и сам читатель.

Пример 4. При исследовании электролитического процесса получения алюминия в планирование эксперимента были включены следующие семь факторов: \tilde{X}_1 — напряжение на электролизере, а; \tilde{X}_2 — время между обработками электролизера, час; \tilde{X}_3 — концентрация фтористого магния в электролите, %; \tilde{X}_4 — концентрация фтористого кальция в электролите, %; \tilde{X}_5 — криолитовое отношение; \tilde{X}_6 — уровень электролита в ванне, см; \tilde{X}_7 — время между операциями смена угольной пены, сутки .

Пример 5. Приведем пример факторов, влияние которых интересовало экспериментатора при оптимизации производства резисторов: \tilde{X}_1 — давление при прессовке, кг/см²; \tilde{X}_2 — температура при прессовке, °С; \tilde{X}_3 — время выдержки под давлением, мин; \tilde{X}_4 — температура в муфеле при прессовке, °С; \tilde{X}_5 — время температурной выдержки, мин; \tilde{X}_6 — дисперсность наполнителя, мк; \tilde{X}_7 — соотношение флюса и наполнителя, г/г; \tilde{X}_8 — давление при шмотировании, кг/см²; \tilde{X}_9 — дисперсность сажи, мк; \tilde{X}_{10} — время выдержки при шмотировании, мин; \tilde{X}_{11} — качество керамики оснований; \tilde{X}_{12} — дисперсность флюсов, мк .

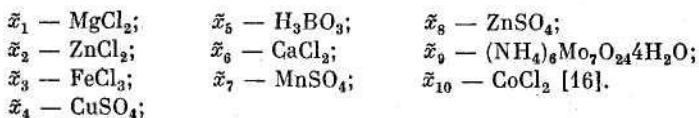
Пример 6. При изучении процесса варки сульфатной целлюлозы в планирование эксперимента были включены такие пять факторов: \tilde{X}_1 — концентрация активной щелочи в варочном растворе (в единицах Na₂O), г/л; \tilde{X}_2 — сульфитность раствора, %; \tilde{X}_3 — конечная температура варки, °С; \tilde{X}_4 — продолжительность подъема температуры до конечной, мин; \tilde{X}_5 — продолжительность варки при конечной температуре, мин.

Пример 7. При оптимизации процесса обогащения молибденовой руды экспериментатор остановил свое внимание на следующих факторах: \tilde{X}_1 — время измельчения руды, мин; \tilde{X}_2 — расход олеата натрия, г/т; \tilde{X}_3 — расход алкилсульфата, г/т; \tilde{X}_4 — расход соды, г/т; \tilde{X}_5 — расход керосина, г/т .

Пример 8. При оптимизации процесса экстракции циркония и гафния из солянокислых растворов в качестве независимых переменных приняты: \tilde{X}_1 — концентрация металла, г/л; \tilde{X}_2 — концентрация кислоты, мол/л; \tilde{X}_3 — концентрация спирта, %; \tilde{X}_4 — соотношение объемов фаз, мл/мл .

Пример 9. В микробиологических исследованиях весьма важной задачей является нахождение оптимального состава питательной

среды. В одной из микробиологических работ проверялось влияние следующих факторов:



Обработка результатов эксперимента велась по пробам, полученным на четвертые сутки выращивания. В экспериментах такого рода возможно также варьировать время выращивания, температуру и т. д.

Пример 10. Для специалистов, занимающихся животноводством, приведем пример факторов, влияющих на откорм свиней.

Определялось соотношение в рационе питательных веществ и стимуляторов, т. е. химических соединений, воздействующих на обмен веществ в организме. Изучалось влияние следующих четырнадцати факторов:

макроэлементы: $\tilde{x}_1 - \text{Ca}$; $\tilde{x}_2 - \text{K}$; $\tilde{x}_3 - \text{Mg}$; $\tilde{x}_4 - \text{Na}$;

микроэлементы: $\tilde{x}_5 - \text{Co}$; $\tilde{x}_6 - \text{Zn}$; $\tilde{x}_7 - \text{Si}$; $\tilde{x}_8 - \text{Fe}$; $\tilde{x}_9 - \text{J}$;
 $\tilde{x}_{10} - \text{Mn}$;

витамины: $\tilde{x}_{11} - \text{A}$; $\tilde{x}_{12} - \text{D}_2$; $\tilde{x}_{13} - \text{B}_{12}$;

антибиотики: $\tilde{x}_{14} - \text{биомицин}$.

Видимо, приведенных примеров достаточно, чтобы можно было составить себе ясное представление о понятии «факторы».

Итак, мы установили, что факторы — это переменные величины, соответствующие способам воздействия внешней среды на объект. **Они определяют как сам объект, так и его состояние.** Требования к факторам: управляемость и однозначность.

Управлять фактором — это значит установить нужное значение и поддерживать его постоянным в течение опыта или менять по заданной программе. В этом достоинство особенность «активного» эксперимента. Планировать эксперимент можно только в том случае, если уровни факторов подчиняются воле экспериментатора.

Факторы должны непосредственно воздействовать на объект исследования. Трудно управлять фактором, если он является функцией других переменных, но в планировании эксперимента могут участвовать сложные факторы, такие, как логарифмы, соотношения и т. д. Факторы должны быть определены операционально.

Требования к совокупности факторов: совместимость и отсутствие линейной корреляции. Выбранное множество факторов должно быть достаточно полным. Если какой-либо существенный фактор пропущен,

это приведет к неправильному определению оптимальных условий или к большой ошибке опыта. Факторы могут быть количественными и качественными.

Точность фиксации факторов должна быть высока. Степень точности определяется диапазоном изменения факторов.

Выбор факторов — очень ответственный этап при подготовке к планированию эксперимента. От удачного выбора зависит успех оптимизации.

После того как вы ознакомились с параметром оптимизации и факторах, можно подойти к выбору модели исследуемого процесса.

3.4. Выбор модели

3.4.1. Модели объектов исследования

При исследовании объектов приходится иметь дело с большим числом объектов (структур и процессов) различной физической природы. Для оптимального управления объектом необходимо иметь формализованное описание этих объектов, что можно осуществить с помощью моделей.

Модель — это формализованное описание основных величин и связей между ними, характерных для рассматриваемого реального объекта. Являясь формой абстрактного описания реальности, модель может быть только приближенной. Поэтому одному и тому же реальному объекту может быть сопоставлено множество моделей, отличающихся друг от друга как по своей структуре и степени детализации свойств рассматриваемого объекта, так и по способу их получения.

Проанализируем, например, поведение газа, заключенного в небольшом сосуде. Если плотность газа не слишком велика, то для его формального описания может быть использована известная модель идеального газа, согласно которой газ представляют в виде совокупности отдельных молекул — шариков бесконечно малого размера, совершающих хаотическое тепловое движение и испытывающих при своем движении абсолютно упругие соударения друг с другом и со стенками окружающего сосуда. Такая модель представляет собой лишь физическую абстракцию, однако на ее основе получено важное с практической точки зрения уравнение состояния газа (уравнение Менделеева—Клапейрона):

$$PV = \frac{M}{\mu} RT, \quad (3.1)$$

где P — давление газа, Па; T — температура газа, К; V — объем, м³; M — масса газа, кг; μ — молярная масса газа, кг/моль; R — $9,31 \cdot 10^3$ Дж/моль — универсальная газовая постоянная.

Уравнение (3.1) может быть названо моделью газа, так как оно представляет собой формализованное описание связей между характерными для газа величинами. *Поскольку модель (3.1) дает количественное описание этих связей, она может быть названа **математической моделью газа**. С точки же зрения **метода получения модель (3.1) является физической**, поскольку она получена в результате применения основных физических законов к молекулам рассматриваемого газа.*

Та же самая модель (3.1) может быть получена на основании обобщения результатов серии экспериментов, в которых для некоторого количества M исследуемого газа осуществлялось варьирование факторов P , V и T . *В этом случае модель (3.1) в соответствии с методом ее получения можно назвать **экспериментально-статистической**.*

Данный пример позволяет сделать вывод о том, что *модели могут классифицироваться как по **форме**, так и по **методу их получения**. Так, модель (3.1) по **форме** является **математической**, а по **методу получения** может быть либо **физической**, либо **экспериментально-статистической**.*

Условимся в дальнейшем вне зависимости от ее формы называть модель *физической*, если она получена как результат применения основных физических или физико-химических законов, определяющих сущность изучаемого объекта, в отличие от статистических моделей, получаемых путем статистического анализа связей между характерными величинами в отрыве от их физической сущности.

Рассмотрим роль и место физических моделей при решении задач исследования и управления объектами. В качестве объекта исследования примем технологически процесс (ТП) изготовления приборов.

Конечной целью исследования является получение численных значений оптимальных параметров объекта исследования, использование которых позволит создать объект исследования с заданными характеристиками (механическими, электрическими, технико-экономическими, надежностными и др). Поэтому при попытке управлять отдельными операциями или всем ТП в целом прежде всего возникает мысль использовать эти характеристики в качестве выходных переменных (параметров). Однако в большинстве случаев такой подход малоэффективен, так как при такой постановке

задачи для осуществления управления требуется построить математическую модель процесса, которая бы связывала **входные (факторы) переменные** ТП с **выходными характеристиками** изготавливаемых приборов. Построить такую модель бывает в большинстве случаев чрезвычайно трудно, а порой и невозможно. Это объясняется тем, что чаще всего отсутствует прямая физическая связь между столь разнородными по своей природе величинами, как, например, количественный состав газовой среды при диффузии и граничная частота усиления транзистора. Поэтому *привлечение физических представлений не дает возможности построить математическую модель ТП. Попытка же использовать аппарат экспериментально-статистического исследования, для которого с формальной точки зрения не требуется знания физического содержания исследуемого процесса, также не приводит к желаемому результату: получаемые при этом модели либо не обладают достаточной точностью предсказания значений выходных параметров приборов, либо имеют приемлемую точность лишь в очень узком диапазоне варьирования технологических факторов — и в том и в другом случае модель оказывается малоприменимой для эффективного управления ТП.*

Большое распространение получил другой подход к решению задачи управления ТП. Он заключается в том, что *связь между технологическими переменными процесса и выходными параметрами изготавливаемых приборов рассматривается не непосредственно, а через структуру прибора, формируемую в результате проведения той или иной технологической операции.* Следовательно, прослеживается логическая цепь **«физика ТП — формируемая структура прибора — выходные параметры прибора»**. При этом исследователю необходимо определить:

а) структуру прибора, которая может обеспечить заданные выходные характеристики;

б) способ (метод) получения данной структуры.

При таком разбиении задачи удается в полной мере использовать весь арсенал представлений как о физике работы прибора (для выяснения связи «структура—характеристики»), так и о физике рассматриваемых ТП (для выяснения связи «ТП—структура»). Можно выделить три группы физических моделей.

Первая группа — это модели, устанавливающие связь между структурой и выходными характеристиками прибора. Модели этой группы получают при рассмотрении физики работы исследуемого прибора. Например, для МДП-транзисторов пороговое напряжение

$$U_{\text{пор}} = \frac{1}{q} (\Phi_3 - \Phi_{\text{п}}) - \frac{Q_{\text{п}}}{\epsilon_{\text{д}}} d + \frac{d}{\epsilon_{\text{д}}} \sqrt{4\epsilon_{\text{п}} q V_{\text{T}} N_{\text{а}} \ln\left(\frac{N_{\text{а}}}{n_i}\right) + 2V_{\text{T}} \ln\left(\frac{N_{\text{а}}}{n_i}\right)}, \quad (3.2)$$

где $\Phi_3 - \Phi_{\text{п}}$ — разность работ выхода затвор—подложка; d — толщина затворного диэлектрика; $\epsilon_{\text{д}}$, $\epsilon_{\text{п}}$ — диэлектрическая проницаемость затворного диэлектрика и подложки; $N_{\text{а}}$ — концентрация акцепторной примеси в подложке (рассматривается n -канальный транзистор); n_i — собственная концентрация носителей заряда в подложке; q — заряд электрона; $V_{\text{T}} = kT/q$ — тепловой потенциал; $Q_{\text{а}}$ — плотность поверхностного заряда на границе затворный диэлектрик—подложка.

С учетом параметров исходных материалов (Φ_3 , $\Phi_{\text{п}}$, $\epsilon_{\text{д}}$, $\epsilon_{\text{п}}$, n_i) выражение (3.2) позволяет определить необходимые параметры формируемой структуры (d , $N_{\text{а}}$, $Q_{\text{п}}$) обеспечивающие требуемое значение $U_{\text{пор}}$ транзистора. При этом учитывают также технологические и физические ограничения, накладываемые на величины d , $N_{\text{а}}$, $Q_{\text{п}}$. Например, для МДП-транзистора со структурой кремний — оксид кремния — металл такими ограничениями являются:

а) $d > 0,05$ мкм, иначе пробивное напряжение затвор—подложка оказывается слишком малым;

б) $N_{\text{а}} < 10^{17}$ см⁻³ в противном случае максимально-допустимое напряжение сток—исток мало из-за пробоя стокового p - n -перехода;

в) $\frac{1}{q} Q_{\text{п}} > 10^{10}$ см⁻², так как меньшей плотности заряда в системе кремний — термический оксид кремния получить не удастся.

Вторая группа физических моделей — это модели, устанавливающие связь между физическими переменными ТП и параметрами получаемой структуры. Модели этой группы используются для управления процессом формирования нужной структуры. Например, толщина l легированного слоя, полученного при диффузии из постоянного источника, связана с температурой T и временем t диффузии соотношением

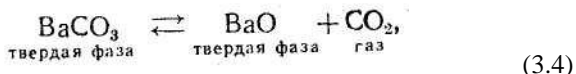
$$l^2 = At e^{-\frac{E_{\text{а}}}{kT}}, \quad (3.3)$$

где A — коэффициент, зависящий от используемого диффузанта и выбранного уровня отсчета концентрации; $k = 1,38 \cdot 10^{-23}$ Дж/град — постоянная Больцмана; E_a — энергия активации процесса диффузии для выбранного диффузанта.

Использование физической модели (3.3) позволяет получить результат, противоречащий обычным представлениям, согласно которым, толщина легированного слоя должна быть прямо пропорциональна времени диффузии.

Третья группа физических моделей — это модели, устанавливающие связь между характерными физическими величинами ТП во время его проведения. Модели этой группы используют для управления ходом ТП.

Рассмотрим в качестве примера процесс разложения карбонатов при проведении активации оксидного катода ЭВП. Процесс описывают с помощью модели гетерогенной химической реакции диссоциации (разложения) BaCO_3 на поверхности катода



для которой закон действующих масс позволяет найти следующее соотношение для равновесного давления единственного газообразного компонента (это давление называют давлением диссоциации рассматриваемого химического соединения):

$$P_{\text{CO}_2} = K_p(T), \quad (3.5)$$

где $K_p(T)$ — константа равновесия, зависящая от температуры и увеличивающаяся с ее ростом.

До тех пор, пока давление диссоциации P_{CO_2} меньше внешнего давления $P_{\text{вн}}$, разложение соединения идет медленно. Равенство $P_{\text{CO}_2} = P_{\text{вн}}$ определяет температуру разложения $T_{\text{разл}}$, начиная с которой наблюдается интенсивное разложение. Таким образом, диссоциацию BaCO_3 можно осуществить двумя способами:

- 1) при $P_{\text{вн}} = \text{const}$ путем увеличения температуры до значения $T_{\text{разл}}$ и выше;
- 2) при $T = \text{const}$ путем уменьшения внешнего давления (путем откачки) до значения $P_{\text{дисс}}(T) = K_p(T)$.

На практике при производстве ЭВП с оксидным катодом для ускорения процесса разложения карбонатов используют одновременно оба способа. Рассмотренные группы физических моделей позволяют охватить все явления и взаимосвязи при производстве ЭП, начиная от задания технологических режимов и количественного состава исходных материалов и кончая выходными параметрами прибора. Однако для управления процессом производства одних только

физических представлений недостаточно, так как физические модели не дают нужной степени детализации — они чаще всего описывают принципиальный вид исследуемых зависимостей между переменными процессами. Поэтому на практике физическую модель используют лишь как основу для постановки эксперимента, в результате которого модель обретает необходимую количественную конкретизацию и становится пригодной для управления.

Рассмотрим в качестве примера использование модели (3.3) для управления процессом диффузии. Согласно модели, толщина диффузионного слоя

$$l = \sqrt{At} e^{-\frac{E_a}{2kT}}.$$

Однако ни коэффициент A , ни энергия активации E_a априори никогда не бывают известны с достаточной точностью. Поэтому ставят серию уточняющих экспериментов, преобразовав модель к виду

$$\ln l = \frac{1}{2} \ln A + \frac{1}{2} \ln t - \frac{E_a}{2k} \frac{1}{T}. \quad (3.6)$$

Если ввести обозначения

$$y = \ln l; \quad x_1 = \ln t; \quad x_2 = 1/T,$$

то модель (3.6) примет вид

$$y = c_0 + c_1 x_1 + c_2 x_2, \quad (3.7)$$

т. е. станет линейной относительно входящих в нее переменных y , x_1 и x_2 .

Коэффициенты c_0 , c_1 , c_2 определяют из обычного двухфакторного эксперимента, для которого требуется провести всего лишь четыре опыта. Коэффициент c_0 позволяет вычислить неизвестный параметр A , а коэффициент c_2 — энергию активации E_a . Коэффициент c_1 , должен быть равен 0,5, однако в результате экспериментов, проведенных в реальных производственных условиях, получают значения, отличающиеся от этой цифры в несколько раз. Аналогично энергия активации может отличаться в несколько раз в зависимости от конкретных условий. Так, в одном из экспериментов по легированию кремниевых пластин бором было получено значение $E_a = 0,6$ эВ, тогда как по известным из литературы данным $E_a = 3,6$ эВ. Такое резкое расхождение свидетельствует о том, что еще мало известно о механизме диффузии через поверхность (диффузия через поверхность идет не так, как в объемном кристалле: например, вблизи поверхности кристалла большую роль может играть диффузия по дислокациям, которая ускоряет проникновение диффузанта в глубь кристалла).

Так как же выбрать модель исследования? Ответить на этот вопрос не так просто. Ниже будут приведены некоторые рекомендации по

выбору модели. Мы говорили, что под моделью понимаем вид функции отклика

$$y = f(x_1, x_2, \dots, x_k).$$

Выбрать модель — значит выбрать вид этой функции, записать ее уравнение. Тогда останется спланировать и провести эксперимент для оценки численных значений констант (коэффициентов) этого уравнения. Но как выбрать модель?

Чтобы постепенно продвигаться к ответу на этот вопрос, необходимо сначала построить геометрический аналог функции отклика — *поверхность отклика*. Будем для наглядности рассматривать случай с двумя факторами.

В случае многих факторов геометрическая наглядность теряется. Мы попадаем в абстрактное многомерное пространство, где у нас нет навыка ориентирования. **В этих случаях необходимо переходить на язык алгебры.** Тем не менее простые примеры, которые будут рассмотрены, окажут помощь при работе с многими факторами.

Требуется изобразить геометрически возможные состояния «черного ящика» с двумя входами. Для этого достаточно располагать плоскостью с обычной декартовой системой координат. По одной оси координат будем откладывать в некотором масштабе значения (уровни) одного фактора, а по другой оси — второго. Тогда каждому состоянию «ящика» будет соответствовать точка на плоскости.

Но, как мы знаем, для факторов существуют области определения. Это значит, что у каждого фактора есть минимальное и максимальное возможные значения, между которыми он может изменяться либо непрерывно, либо дискретно. Если факторы совместимы, то границы образуют на плоскости некоторый прямоугольник, внутри которого лежат точки, соответствующие состояниям «черного ящика». Пунктирными линиями на рис. 3.11 обозначены границы областей определения каждого из факторов, а сплошными — границы их совместной области определения.

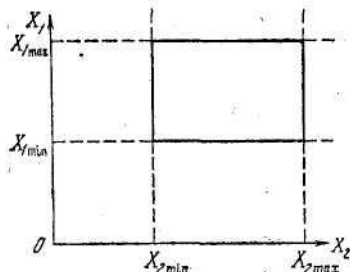


Рис. 3.11. Область определения факторов

Чтобы указать значение параметра оптимизации, требуется еще одна ось координат. Если ее построить, то поверхность отклика будет выглядеть так, как на рис. 3.12.

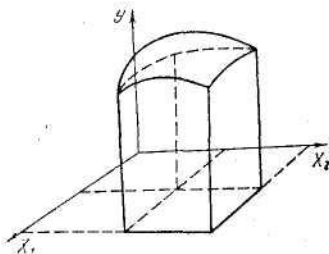


Рис. 3.12. Поверхность отклика

Пространство, в котором строится поверхность отклика, мы будем называть **факторным пространством**. Оно задается координатными осями, по которым откладываются значения факторов и параметра оптимизации (иногда под факторным пространством понимается пространство, образованное только осями факторов). Размерность факторного пространства зависит от числа факторов. При многих факторах поверхность отклика уже нельзя изобразить наглядно и приходится ограничиваться только алгебраическим языком.

Но для двух факторов можно даже не переходить к трехмерному пространству, а ограничиться плоскостью. Для этого достаточно произвести сечение поверхности отклика плоскостями, параллельными плоскости X_1OX_2 , и полученные в сечениях линии спроектировать на эту плоскость. Так строят, например, изображения гор и морских впадин на географических картах (рис. 3.13).

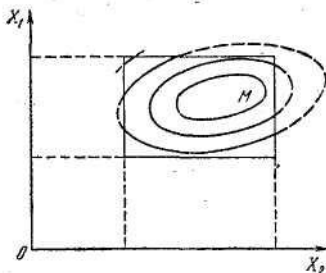


Рис. 3.13. Проекция сечений поверхности отклика на плоскость

Точка *M* на рисунке — это и есть та оптимальная точка, которую мы ищем. Каждая, линия соответствует постоянному значению параметра оптимизации. Такая линия называется *линией равного отклика*. Существует соответствие между состоянием «ящика» и значением параметра оптимизации: **каждому возможному состоянию «ящика» соответствует одно значение параметра оптимизации**. Однако обратное неверно: одному возможному значению параметра оптимизации может соответствовать и одно, и несколько, и сколько угодно состояний «ящиков».

Эти утверждения справедливы, если не учитывать ошибок в определении значений параметра оптимизации. К вопросу об оценке и учете этих ошибок мы вернемся ниже, а пока не будем принимать их во внимание.

Теперь, когда мы можем представить себе поверхность отклика, необходимо вернуться к основному вопросу: как ставить эксперимент, чтобы найти оптимум при минимуме затрат? Это прежде всего вопрос стратегии.

Если бы мы располагали таблицей, в которой содержались бы все возможные состояния объекта и соответствующие им отклики, то особой необходимости в построении математической модели не было бы. Просто мы бы выбрали то (или те) состояние, которое соответствует наилучшему отклику. Но мы знаем, что перебор возможных состояний очень велик, и поэтому отказываемся от практической реализации этой возможности.

Другая возможность — случайный выбор некоторого числа состояний и определение откликов в них, в надежде, что среди этих состояний попадутся оптимальное или по крайней мере близкие к нему состояния.

Третья возможность — строить математическую модель, чтобы с ее помощью предсказывать значения откликов в тех состояниях, которые не изучались экспериментально. Если не можем измерить отклик в каждом состоянии, то сумеем хоть предсказывать результат. Причем даже не в каждом состоянии, а только в наиболее интересных, в тех, которые приближают нас к оптимуму.

Такая стратегия приводит нас к шаговому принципу, лежащему в основе рассматриваемого метода планирования эксперимента.

3.4.2. Шаговый принцип

За отказ от полного перебора состояний надо чем-то платить. Цена — это предположения, которые мы должны сделать относительно

свойств неизвестной нам модели до начала эксперимента (как говорят, *априори*).

Некоторые из предположений мы никогда не сможем проверить. Такие предположения называются постулатами. Если в действительности они не выполняются, то весьма возможно, что мы не найдем оптимум. Точнее, мы примем за оптимум то, что на самом деле им не является (хотя, быть может, нас и удовлетворит).

Какие же предположения о свойствах поверхности отклика мы делаем? **Главное** — это *непрерывность поверхности, ее гладкость и наличие единственного оптимума* (быть может, и на границе области определения факторов).

Эти постулаты позволяют представить изучаемую функцию в виде степенного ряда в окрестности любой возможной точки факторного пространства (такие функции в математике называются аналитическими). Кроме того, если мы придумаем какой-то способ постепенного приближения к оптимальной точке, нужно, чтобы результат не зависел от исходной точки. Если оптимум один, то неважно, приближаемся мы к нему справа или слева, а если их несколько, да они еще неравноценны.

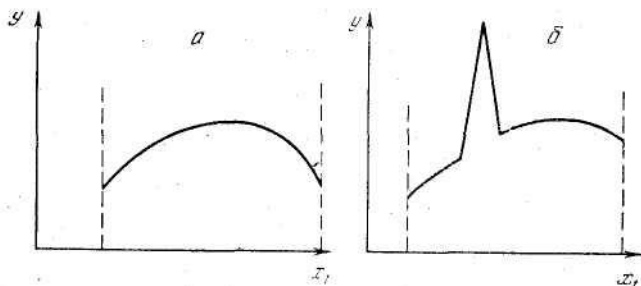


Рис. 3.14. Примеры функций отклика для одного фактора

На рис. 3.14 приведены две картинки, изображающие функции отклика для одного фактора. На рис. 14, *a* показан благоприятный случай. На рис. 14, *b* — много нарушений. Здесь и два экстремума (оптимума) и пик (нарушение гладкости и непрерывности).

Если в поисках оптимума мы начнем последовательно двигаться слева направо, то найдем наименьший из максимумов и вряд ли узнаем о существовании второго, наибольшего. Правда, он так локализован и остер, что его возможно пропустить и при движении с правого конца, если ставить опыты не во всех точках.

Следует обратить внимание на то, что требование непрерывности не согласуется с представлением о дискретных уровнях факторов. Однако

в действительности это не страшно. Мы можем считать, что фактор принимает непрерывный ряд значений (если даже некоторые значения не имеют смысла или физически нереализуемы). Важно только помнить о таком соглашении при использовании результатов. А для построения математической модели это создает значительные удобства.

Так как мы заранее считаем, что предпосылки выполняются, то надо максимально использовать возможности, которые при этом открываются.

Если, например, мы будем знать значения параметра оптимизации в нескольких соседних точках факторного пространства, мы сможем (в силу гладкости и непрерывности функции отклика) представить себе результаты, которые можно ожидать в других соседних точках. Следовательно, можно найти такие точки, для которых ожидается наибольшее увеличение (или уменьшение, если мы ищем минимум) параметра оптимизации. Тогда ясно, что следующий эксперимент надо переносить именно в эти точки. Надо продвигаться в этом направлении, пренебрегая остальными. Сделав новый эксперимент, снова можно оценить направление, в котором скорее всего следует двигаться. В силу единственности оптимума мы, таким образом, рано или поздно непременно его достигнем. *Это и есть шаговый принцип.*

Сделаем некоторые пояснения. Мы выбираем в факторном пространстве какую-то точку и рассматриваем множество точек в ее окрестности, т. е. выбираем в области определения факторов малую подобласть. Здесь мы хотим провести эксперимент, на основании которого должна быть построена первая модель. Эту модель мы намерены использовать для предсказания результатов опытов в тех точках, которые не входили в эксперимент. Если эти *точки лежат внутри нашей подобласти, то такое предсказание называется интерполяцией, а если вне — экстраполяцией.* Чем дальше от области эксперимента лежит точка, для которой мы хотим предсказать результат, тем с меньшей уверенностью это можно делать. Поэтому мы вынуждены экстраполировать недалеко и использовать результаты экстраполяции для выбора условий проведения следующего эксперимента. Дальше цикл повторяется.

Попутно полученную модель можно использовать для проверки различных гипотез о механизме изучаемого явления или о его отдельных сторонах. Например, если вы предполагаете, что увеличение значения некоторого фактора должно приводить к увеличению значения параметра оптимизации, то с помощью модели

можно узнать, так ли это. Такая проверка называется *интерпретацией модели*.

На рис. 3.15 изображены два варианта поиска оптимума для одной и той же поверхности.

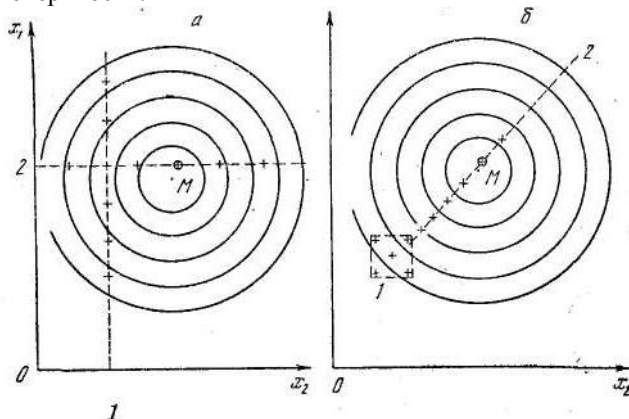


Рис. 3.15. Два способа поиска оптимума

Крестиками на рисунке обозначены условия опытов. В случае *a* использован подход, который иногда называют классическим (метод Гаусса—Зейделя). Он состоит в том, что сначала последовательно изменяются значения одного фактора. (На рисунке этот эксперимент обозначен 1.) Затем находится и фиксируется наилучшее значение этого фактора. В этих условиях последовательно изменяются значения второго фактора (2) и т. д. (если больше факторов).

В случае *б* представлен простейший вариант шаговой процедуры. Сначала изучается локальная область (1), затем определяется наиболее интересное направление и в этом направлении ставятся следующие опыты (2).

Оказалось (см. рис. 3.15), что в обоих случаях достигнут одинаковый результат при одинаковом суммарном количестве опытов.

Возникает вопрос, всегда ли эти две процедуры эквивалентны? Что требуется? Выяснить, нет ли нарушений наших предпосылок. Легче всего установить, сколько оптимумов (экстремумов) имеет изображенная функция. Если экстремумов больше одного, то уже нарушена предпосылка. Кроме того, существенно, нет ли каких-нибудь нарушений гладкости и непрерывности функции (например, пиков).

Эффективность зависит от вида поверхности, а также от того, в какой последовательности перебираются факторы в случае *a* и из окрестностей какой точки начат эксперимент в случае *б*.

Если вместо окружностей, которые задают линии равных откликов, нарисовать эллипсы, главные оси которых составляют некоторый острый угол с осями координат, то можно увидеть, что эффективность процедуры окажется различной.

Иллюстрация, которая показывает правильность вашего ответа, приведена на рис. 3.16. Однако в жизни не всегда удается за один цикл достигнуть оптимума.

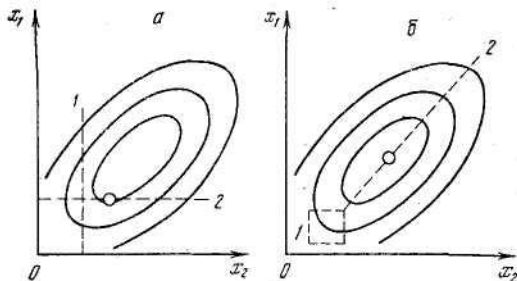


Рис. 3.16. Два способа поиска оптимума.

Но несомненно, что по крайней мере в отношении результата процедура *б*, т. е. шаговый метод, в среднем эффективнее, чем процедура *а*. Можно придумать и более конкурентноспособные процедуры, чем *а*, но они обычно требуют значительно больше опытов.

Теперь займемся выбором модели для первого эксперимента.

3.4.3. Полиномиальные модели

Модели бывают разные. Моделей бывает много. Чтобы выбрать одну из них, надо понять, что мы хотим от модели, какие требования мы к ней предъявляем. Теперь мы сможем сформулировать эти требования.

Исходя из выбранной стратегии, ясно, что **главное требование к модели** — это способность предсказывать направление дальнейших опытов, причем предсказывать с требуемой точностью. Так как до получения модели мы не знаем, какое направление нам понадобится, то естественно требовать, чтобы точность предсказания во всех возможных направлениях была одинакова.

Это значит, что в некоторой подобласти, в которую входят и координаты выполненных опытов, предсказанное с помощью модели значение отклика не должно отличаться от фактического больше чем на некоторую заранее заданную величину. *Модель, которая удовлетворяет такому или какому-либо аналогичному требованию,*

называется **адекватной**. Проверка выполнимости этого требования называется проверкой **адекватности модели**. Разработаны специальные статистические методы, с помощью которых проверяется адекватность. Мы их рассмотрим позже.

Если несколько различных моделей отвечают нужным требованиям, то следует предпочесть ту из них, которая является самой простой. На рис. 3.17 изображена логарифмическая функция.

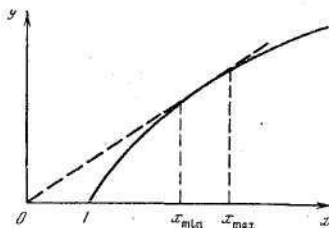


Рис. 3.17. График логарифмической функции

На некотором отрезке $[x_{\min}, x_{\max}]$ она с удовлетворительной точностью описывается двумя уравнениями:

$$y = \log_b x, \quad (3.8)$$

$$y = bx. \quad (3.9)$$

В уравнении (3.9) b — коэффициент, который мы можем оценить, например, по результатам эксперимента. Какое из уравнений, (3.8) или (3.9), проще?

Простота — вещь относительная. Если вы заранее не сформулируете точно, что называется простым, а что сложным, то невозможно произвести выбор. Вот почему на наш вопрос не было никакого другого ответа, кроме «не знаю».

На будущее мы договоримся, что *при прочих равных условиях мы всегда будем предпочитать степенные ряды*. Точнее, *отрезки степенных рядов — алгебраические полиномы*. При таком соглашении можно сказать, что уравнение (3.9) проще, чем уравнение (3.8).

Фактически мы произвели выбор класса моделей. Мы сказали, что всегда, когда это возможно, будем искать модель среди полиномов. Построение полинома возможно в окрестностях любой точки факторного пространства, поскольку мы предположили, что функция является аналитической.

Выбрать — значит сравнить. А как сравнить между собой классы моделей, если свойства объекта заранее неизвестны? Остается предполагать, что нам будут редко встречаться задачи, в которых

исходные постулаты окажутся существенно неверными. Если это так, то мы действительно выбрали наиболее простой, удобный и математически разработанный класс моделей.

Выпишем полиномы для случая двух факторов. Они будут различаться по максимальным степеням входящих в них переменных.

Полином нулевой степени: $y = b_0$.

Полином первой степени: $y = b_0 + b_1x_1 + b_2x_2$.

Полином второй степени:

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2.$$

Полином третьей степени:

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{112}x_1^2x_2 + b_{112}x_1x_2^2 + b_{111}x_1^3 + b_{222}x_2^3.$$

Итак, мы представили неизвестную нам функцию отклика полиномом. Операция замены одной функции другой, в каком-то смысле эквивалентной функцией называется аппроксимацией. Значит, мы аппроксимировали неизвестную функцию полиномом.

Но полиномы бывают разных степеней. Какой взять на первом шаге?

Эксперимент нужен только для того, чтобы найти численные значения коэффициентов полинома. Поэтому чем больше коэффициентов, тем больше опытов окажется необходимым. А мы стремимся сократить их число. Значит, надо найти такой полином, который *содержит как можно меньше коэффициентов, но удовлетворяет требованиям, предъявленным к модели. Чем ниже степень полинома при заданном числе факторов, тем меньше в нем коэффициентов.*

Мы хотим, чтобы модель хорошо предсказывала направление наискорейшего улучшения параметра оптимизации. Такое направление называется направлением градиента. Ясно, что движение в этом направлении приведет к успеху быстрее, чем движение в любом другом направлении (это значит, что будет достигнута экономия числа опытов).

Как вы думаете, можно ли в этой связи всегда использовать полином первой степени?

С одной стороны, он содержит информацию о направлении градиента, с другой — в нем минимально возможное число коэффициентов при данном числе факторов. Единственное опасение в том, что неясно, будет ли линейная модель всегда адекватной. Ответ зависит еще и от объекта.

Вопрос в том, как выбрать подобласть в факторном пространстве, чтобы линейная модель оказалась адекватной. Условие аналитичности функции отклика гарантирует нам эту возможность. Всегда существует такая окрестность любой точки (точнее, почти любой точки), в которой

линейная модель адекватна. Размер такой области заранее не известен, но адекватность можно проверять по результатам эксперимента. Значит, выбрав сначала произвольную подобласть, мы, рано или поздно, найдем ее требуемые размеры. И как только это случится, воспользуемся движением по градиенту.

На следующем этапе мы будем искать линейную модель уже в другой подобласти. Цикл повторяется до тех пор, пока движение по градиенту не перестанет давать эффект. Это значит, что мы попали в область, близкую к оптимуму. Такая область называется «почти стационарной». Здесь линейная модель уже не нужна. Либо попаданием в почти стационарную область задача решена, либо надо переходить к полиномам более высоких степеней, например второй степени, чтобы подробнее описать область оптимума.

Удачный выбор подобласти имеет, как видно, большое значение для успеха всей работы. Он связан с интуитивными решениями, которые принимает экспериментатор на каждом этапе. Как это делается, мы рассмотрим ниже.

Кроме задачи оптимизации, иногда возникает задача построения интерполяционной модели. В этом случае нас не интересует оптимум. Просто мы хотим предсказывать результат с требуемой точностью во всех точках некоторой заранее заданной области. Тут не приходится выбирать подобласть. Необходимо последовательно увеличивать степень полинома до тех пор, пока модель не окажется адекватной. Если адекватной оказывается линейная или неполная квадратная модель (без членов, содержащих квадраты факторов), то ее построение аналогично тому, что требуется для оптимизации.

Подведем некоторые итоги выбора модели.

Как мы уже говорили количество факторов и характер их взаимосвязей с выходной переменной определяют сложность объекта исследования. При наличии качественной статистической информации о факторах и зависящей от них выходной переменной можно построить математическую модель объекта исследования или функцию отклика $y=f(x_1, x_2, \dots, x_n)$, связывающую параметр оптимизации с факторами, которые варьируются при проведении опытов.

Пространство с координатами x_1, x_2, \dots, x_n принято называть факторным, а графическое изображение функции отклика в факторном пространстве — поверхностью отклика.

При описании объектов, находящихся в стационарном состоянии, математическая модель чаще всего представляется полиномом:

$$y = f(x_1, x_2, \dots, x_n, \beta_0, \beta_1, \dots, \beta_n). \quad (3.10)$$

Поскольку в реальном процессе всегда существуют неуправляемые и неконтролируемые переменные, величина y носит случайный характер. Поэтому при обработке экспериментальных данных получаются так называемые выборочные коэффициенты регрессии $b_0, b_1, \dots, b_i, \dots, b_n$, являющиеся оценками теоретических коэффициентов $\beta_0, \beta_1, \dots, \beta_i, \dots, \beta_n$.

Тогда математическая модель в форме уравнения регрессии в общем случае будет иметь вид

$$\hat{y} = b_0 + \sum_i b_i x_i + \sum_{i < j} b_{ij} x_i x_j + \sum_i b_{ii} x_i^2 + \dots \quad (3.11)$$

Если анализируются нестационарные, т. е. изменяющиеся во времени состояния объекта, приходится рассматривать не случайные величины, как ранее, а случайные процессы. **Случайный процесс можно рассматривать как систему, состоящую из бесконечного множества случайных величин.** При моделировании таких объектов использовать модель в виде (3.11) уже недопустимо — необходимо переходить к специальным **интегрально-дифференциальным моделям.**

В отличие от статистических математических моделей существуют физико-химические модели (иногда их называют аналитическими). При построении таких моделей используются физико-химические закономерности моделируемых процессов. Построение физико-химических моделей сопряжено с проведением обширных и длительных исследований, поскольку необходимо рассматривать микроструктуру процессов (чаще это иерархия процессов) и описывать их математически. **Модели в этом случае представляются системами алгебраических и различных дифференциальных уравнений.** Не всегда удастся рассчитать численные значения коэффициентов, входящих в эти системы, и тогда опять прибегают к эксперименту. Используя теорию случайных процессов, удастся найти численные значения коэффициентов модели по данным статистической информации. Эта задача известна в теории как **задача идентификации.**

Составлению плана эксперимента всегда должны предшествовать сбор априорной информации для составления характеристики объекта исследования, опыты по наладке экспериментальной установки и при необходимости — опыты для установления области определения наиболее существенных факторов и выходной переменной.

Но вернемся к вопросу выбора модели

Итак, мы выбрали модель, которую будем систематически использовать на первом этапе планирования эксперимента. Это алгебраический полином первой степени — линейная модель.

Чтобы произвести такой выбор, нам понадобилось научиться изображать поверхность отклика в факторном пространстве, задаваемом прямоугольными декартовыми координатами, по осям которых откладываются в некотором масштабе значения (уровни) факторов и значения параметров оптимизации. Поверхность отклика задана только в совместной области определения факторов. В этой области каждому возможному набору значений факторов (состоянию объекта) соответствует единственное значение параметра оптимизации. Для уменьшения размерности факторного пространства при геометрическом построении поверхности отклика можно использовать сечения.

Мы выяснили, что математическая модель требуется для предсказания направления градиента, т. е. направления, в котором величина параметра оптимизации улучшается быстрее, чем в любом другом направлении. Такая модель позволяет избежать полного перебора состояний объекта и тем самым уменьшить количество опытов, необходимых для отыскания оптимума.

Отказ от полного перебора требует оплаты в виде предположений о свойствах поверхности отклика, которые мы не сможем проверить. Такие предположения можно выбрать по-разному. Мы выбрали предположения об аналитичности функции отклика и об единственности оптимума. Аналитической называется такая функция, которую можно разложить в степенной ряд в окрестностях любой точки из области ее определения.

Используя эти предпосылки, можно предложить процедуру поиска оптимума, основанную на шаговом принципе. Этот принцип гласит: проводи короткие (насколько возможно) серии опытов, по их результатам строй математическую модель, используй модель для оценки градиента, ставь новые опыты только в этом направлении. Получается циклический процесс, который заканчивается при попадании в область, близкую к оптимуму («почти стационарную» область).

Чтобы выбрать теперь конкретную модель, надо сформулировать конкретные требования. К ним относятся - адекватность и простота.

Под *адекватностью* понимается способность модели предсказывать результаты эксперимента в некоторой области с требуемой точностью. После реализации опытов можно проверить адекватность модели.

Простота — вещь относительная. Мы просто условились считать алгебраические полиномы самыми простыми. Это соглашение базируется на накопленном разными исследователями опыте работы с такими моделями и обычно удовлетворяет экспериментатора. Кроме того, полиномом линейен относительно неизвестных коэффициентов, что упрощает обработку наблюдений.

Так мы выбрали класс моделей. Осталось выбрать степень полинома и подобласть, в которой надо начинать эксперимент. Эти выборы связаны между собой. Однако важно, что в принципе возможен такой выбор области, при котором линейная модель окажется адекватной. Этого достаточно, чтобы оценить градиент.

Выбор области связан с теми интуитивными решениями, которые принимает экспериментатор на каждом этапе работы.

Попутно мы упомянули о задаче построения интерполяционных моделей, которые используются для предсказания откликов во всей области. Область фиксируется заранее. ***Надо последовательно повышать степень полинома до тех пор, пока не найдется адекватная модель.***

Подведем некоторые итоги этого раздела.

Теорией и практикой эксперимента выработаны определенные требования (условия), которым должны удовлетворять независимые (факторы) и зависимые (параметры) переменные. Поэтому на стадии подготовки к проведению эксперимента весьма полезны приведенные ниже рекомендации.

1. При выборе выходной переменной необходимо учитывать, что она должна иметь количественную характеристику, т. е. должна измеряться; однозначно оценивать (измерять) работоспособность объекта исследования; быть статистически эффективной, т. е. иметь возможно меньшую дисперсию при проведении опытов (это позволяет четко различать опыты); отражать как можно более широкий спектр исследуемого явления, т. е. обладать универсальностью (практически это требование обеспечить трудно, тогда рекомендуют пользоваться так называемой обобщенной переменной); иметь достаточно четкий физический смысл.

Удачный выбор выходной переменной в значительной степени определяется уровнем знания объекта.

2. При выборе факторов нужно выполнять следующие требования: фактор должен быть регулируемым, т. е. определенным регулирующим устройством фактор должен изменяться от значения x'_i до значения x''_i (например, расход вещества может изменяться от 30 до 40 л/ч или количество вещества A в смеси от 10 до 80%, это свойство

называют операционной определенностью); точность измерения и управления фактором должна быть известна и достаточно высока (хотя бы на порядок выше точности измерения выходной переменной) очевидно, что низкая точность измерения фактора уменьшает возможности воспроизведения эксперимента; связь между факторами должна быть как можно меньшей (в пределе должна отсутствовать), это свойство называют однозначностью факторов, что соответствует независимости одного фактора от другого.

Ряд требований предъявляется одновременно к факторам и выходной переменной: факторы и выходная переменная должны иметь области определения, заданные технологическими или принципиальными ограничениями (пример технологического ограничения — максимальная производительность компрессора, подающего газ в реактор; пример принципиального — температура кристаллизации жидкого продукта, образующегося в результате реакции); области определения факторов должны быть таковы, чтобы при их предельных значениях значение выходной переменной оставалось в своих границах; между факторами и выходной переменной должно существовать однозначное соответствие (причинно-следственная связь).

Успех экспериментирования в значительной мере обязан теории эксперимента, которая признана дать экспериментатору ответы на следующие вопросы:

1. Как нужно организовать эксперимент, чтобы наилучшим образом решить поставленную задачу (в смысле затрат времени, средств или точности результатов)?
2. Как следует обрабатывать результаты эксперимента, чтобы получить максимальное количество информации об исследуемом объекте?
3. Какие обоснованные выводы можно сделать об исследуемом объекте по результатам эксперимента?

Как уже упоминалось, основой теории эксперимента является статистическое представление об эксперименте (рассматриваются случайные величины или случайные функции). Это представление отвечает действительности: как правило, итоги эксперимента связаны с некоторой неопределенностью, получающейся в результате влияния неконтролируемых факторов, случайного характера процесса на микроуровне, изменений условий эксперимента, ошибок измерения и др.

Статистическое представление об эксперименте позволило перейти к многофакторному активному эксперименту, в котором удастся

надежно отделить влияние факторов от шумового фона и перейти к статистически обоснованным методам анализа результатов и принятия решений.

Теория эксперимента использует принципы и концепции кибернетики, такие как математическое моделирование, оптимизация. Большое значение имеет концепция оптимального использования факторного пространства с точки зрения определенных характеристик точности решения. Так, удается получать оценки коэффициентов модели, обладающие минимальной дисперсией, или удается получать наилучшие предсказательные свойства модели.

Теория эксперимента указывает исследователю точную логическую схему и способы поиска решения задач на разных этапах исследования. Можно представить весь процесс исследования циклами, повторяющимися после решения каждой из последовательных задач исследования, причем объем знаний об объекте непрерывно увеличивается.

4. Основные методы планирования экстремального эксперимента

4.1. Сущность теории планирования эксперимента

Традиционно измерение параметров как отдельных элементов и узлов систем, так и систем в целом осуществляется последовательным определением исследуемых зависимостей при изменении одной величины и постоянстве остальных. Так, изучение функции двух переменных $y = \varphi(x_1, x_2)$ проводится сначала при фиксированном x_2 (при этом получают $y = \varphi(x_1 |_{x_2=const})$), а затем при фиксированном x_1 (измеряют $y = \varphi(x_2 |_{x_1=const})$). Оба вида кривых представляют собой сечения поверхности $y = \varphi(x_1, x_2)$ плоскостями, параллельными координатным плоскостям $yo x_1$ и $yo x_2$. Например, при определении вида амплитудной характеристики любого устройства в зависимости от частоты изменяют амплитуду входного гармонического сигнала на определенной фиксированной частоте и регистрируют значение выходного сигнала. Затем измерения повторяют на другой частоте и т. д.

В настоящее время широко используются методы математического планирования эксперимента. Если в сравнительно простых измерениях с достаточной для практики точностью можно пользоваться упомянутыми традиционными методами, то серьезные экспериментальные

исследования с многими изменяющимися параметрами и входными величинами без планирования эксперимента просто невозможны.

Рассмотрим простой пример измерения величины сопротивления каждого из трех последовательно включенных резисторов $R1, R2, R3$. Упомянутыми традиционными методами измеряют сопротивления каждое в отдельности и, предварительно выполнив еще одно измерение, с замкнутыми накоротко выводами измерительного прибора для установки нуля его показаний. Схему всего эксперимента, состоящего из четырех опытов, можно представить матрицей планирования X , столбцы которой отражают операцию измерения сопротивлений $R1, R2, R3$ и для общности математический выражений обозначены x_1, x_2, x_3 ; зависимая переменная представлена вектором Y :

$$X = \begin{matrix} & \begin{matrix} x_1 & x_2 & x_3 \end{matrix} \\ \begin{matrix} - \\ + \\ - \\ - \end{matrix} & \begin{bmatrix} -1 & -1 & -1 \\ +1 & -1 & -1 \\ -1 & +1 & -1 \\ -1 & -1 & +1 \end{bmatrix} \end{matrix}, \quad Y = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad (4.1)$$

где знак «+» указывает, что резистор подключен к измерительному прибору, а знак «—» означает, что значение сопротивления резистора не измеряется. Значение сопротивления оценивается по результатам двух измерений; установки нуля прибора и непосредственного измерения величины сопротивления. При этом получаем следующие величины сопротивлений:

$$R1 = y_1 - y_0; \quad R2 = y_2 - y_0; \quad R3 = y_3 - y_0.$$

Дисперсия результатов измерений имеет вид

$$\sigma^2 \{R1\} = \sigma^2 \{y_1 - y_0\}; \quad \sigma^2 \{R2\} = \sigma^2 \{y_2 - y_0\}; \\ \sigma^2 \{R3\} = \sigma^2 \{y_3 - y_0\}; \quad \sigma^2 \{R1\} = \sigma^2 \{R2\} = \sigma^2 \{R3\} = 2 \sigma_y^2.$$

где σ_y — среднеквадратическая ошибка.

Измерение этих же последовательно включен и сопротивлений можно провести по-другому в соответствии со следующей матрицей;

$$X = \begin{matrix} & \begin{matrix} x_1 & x_2 & x_3 \end{matrix} \\ \begin{matrix} + \\ - \\ - \\ + \end{matrix} & \begin{bmatrix} +1 & -1 & -1 \\ -1 & +1 & -1 \\ -1 & -1 & +1 \\ +1 & +1 & +1 \end{bmatrix} \end{matrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}. \quad (4.2)$$

Как и в предыдущем случае, измеряется каждое из сопротивлений в отдельности, а в последнем четвертом опыте измеряется полное

сопротивление всех трех резисторов, предварительная установка нуля прибора не производится. Значение каждого из сопротивлений получаем перемножением элементов вектора Y на элементы столбцов x_1, x_2, x_3 матрицы X с делением результатов на два, так как измерение каждого из сопротивлений выполнено дважды. Таким образом, находим:

$$\begin{aligned} R1 &= 0,5 (y_1 - y_2 - y_3 + y_4); \\ R2 &= 0,5 (-y_1 + y_2 - y_3 + y_4); \\ R3 &= 0,5 (-y_1 - y_2 + y_3 + y_4). \end{aligned}$$

Дисперсия результатов измерения в этом случае

$$\begin{aligned} \sigma^2 \{R1\} &= \sigma^2 [0,5 (y_1 - y_2 - y_3 + y_4)] = 0,25\sigma_y^2 \cdot 4 = \sigma_y^2; \\ \sigma^2 \{R2\} &= \sigma_y^2; \quad \sigma^2 \{R3\} = \sigma_y^2. \end{aligned}$$

Таким образом, новая схема измерений дает уменьшение дисперсии вдвое по сравнению с традиционным методом при одинаковом количестве измерений. Чтобы получить такую же точность традиционным методом, необходимо количество измерений увеличить вдвое. Повышение точности получено благодаря тому, что значение каждого сопротивления по второй схеме измерялась во всех четырех опытах, а по первой схеме — только в двух. Рассмотренный пример показывает, что даже сравнительно простые эксперименты могут быть спланированы лучше, чем они выполнялись традиционно. С усложнением экспериментов эффективность их планирования возрастает. Объектом исследования с применением теории планирования экспериментов могут быть любые объекты.

Рассмотрим основные определения, сущность и задачи планирования экспериментов.

Входные величины объектов исследования могут качественно отличаться друг от друга, например, амплитуда и частота сигнала, входная и выходная проводимости, стабильность питающего напряжения и т. д., поэтому в теории планирования эксперимента входные параметры Z_1, Z_2, \dots, Z_n принято именовать, как мы знаем, общим названием факторы. В целях упрощения обозначений и вычислений целесообразно пользоваться нормированными переменными:

$$\begin{aligned} x_i &= (Z_i - Z_{0i})/\Delta Z_i; \quad \Delta Z_i = 0,5 (Z_{i \max} - Z_{i \min}); \\ Z_{0i} &= 0,5 (Z_{i \max} + Z_{i \min}). \end{aligned}$$

В этих соотношениях $Z_{i \max}$ соответствует $x_i = +1$, а $Z_{i \min}$ соответственно $x_i = -1$.

Выходные величины y_1, y_2, \dots, y_n также могут бы качественно различными, например, передаточная функция, уровень выходного сигнала или его спектральных составляющих, помехоустойчивость, надежность и т. д.— они получили, как мы знаем, название отклик (функция цели, параметр оптимизации).

Как нам известно, модель объекта представляет собой аналитическую зависимость отклика от факторов. Чаще всего эта зависимость неизвестна, известными являются факторы x_i , и выходные величины отклика y_i . Чаще встречается задача исследования одной выходной величины y как функции нескольких факторов-

$$y = \varphi(x_1, x_2, \dots, x_k). \quad (4.3)$$

Вид этой зависимости определяется из физической сущности, а параметры вычисляются в соответствии с результатами эксперимента. Поэтому модель называют также эмпирической. Следует отметить, что модель объекта может быть иногда построена исходя из теоретически обоснованного понимания сущности процессов, происходящих в объекте, например, протекания тока в цепях и активных элементах, выделения спектральных составляющих, образования фазовых сдвигов между входными и выходными переменными и т. д. *Созданную таким образом модель называют теоретической.* Например, работу преобразователя частоты можно описать как выделение отклика суммарной или разностной частоты гармонических составляющих входных факторов при их перемножении.

Планирование эксперимента позволяет решать следующие задачи: исследование механизма физического явления, отыскание экстремума отклика, определение модели объекта и т. п. . Исследовать механизм явления означает определить аналитическое выражение

$$y = f(x_1, x_2, \dots, x_k), \quad (4.4)$$

которое достаточно точно описывало бы неизвестную зависимость (4.3) в пределах области возможных значений факторов x_i , называемую область определения факторов, или факторным пространством. Область определения двух факторов x_1, x_2 представляет собой часть плоскости $x_1 \text{O} x_2$ декартовой системы координат $x_1 \text{O} x_2 y$ и называется двухфакторным пространством, а эксперимент — двухфакторным экспериментом. Могут быть также одно- и многофакторные эксперименты. Эксперименты, направленные на раскрытие механизма исследуемого явления и определяющие аналитическую зависимость (4.4), называют также интерполяционными или регрессионными. Эксперименты, позволяющие находить экстремум отклика в области его определения, называют экстремальными.

Все выше сказанное имеет непосредственное отношение к планированию эксперимента, но прежде чем приступить к его планированию, следует дать ответы на вопросы, поставленные в разделе 3. Прежде всего нас интересует, как выбрать локальную область факторного пространства, где ее выбирать и какого размера она должна быть? Это важный этап принятия неформализованных решений, предшествующих построению плана первой серии эксперимента.

Здесь мы впервые сталкиваемся с проблемой принятия решений при планировании эксперимента. Далее мы уже не расстанемся с этой темой. Поэтому остановимся на особенностях этих этапов решения задачи. Весь процесс исследования можно считать состоящим из последовательности этапов, часть из которых полностью формализована, а часть требует «интуитивных» решений. Причем, по мере развития теории формальные этапы будут играть все большую роль, но до конца не вытеснят неформализованные этапы.

4.2. Принятие решений перед планированием эксперимента

При выборе области эксперимента прежде всего надо *оценить границы областей определения факторов*. При этом должны учитываться ограничения нескольких типов.

Первый тип — принципиальные ограничения для значений факторов, которые не могут быть нарушены ни при каких обстоятельствах. Например, если фактор — температура, то нижним пределом будет абсолютный нуль.

Второй тип — ограничения, связанные с технико-экономическими соображениями, например, со стоимостью сырья, дефицитностью отдельных компонентов, временем ведения процесса.

Третий тип ограничений, с которым чаще всего приходится иметь дело, определяется конкретными условиями проведения процесса, например, существующей аппаратурой, технологией, организацией. В реакторе, изготовленном из некоторого материала, температуру нельзя поднять выше температуры плавления этого материала или выше рабочей температуры данного катализатора.

Оптимизация обычно начинается в условиях, когда объект уже подвергался некоторым исследованиям. *Информацию, содержащуюся в результатах предыдущих исследований, будем называть априорной* (т. е. *полученной до начала эксперимента*). Мы можем использовать априорную информацию для представления представления о параметре оптимизации, о факторах, о наилучших условиях ведения процесса и характере поверхности отклика, т. е. о том, как сильно меняется

параметр оптимизации при небольших изменениях значений факторов, а также о кривизне поверхности. Для этого можно использовать графики (или таблицы) однофакторных экспериментов, осуществившихся в предыдущих исследованиях или описанных в литературе. Если однофакторную зависимость нельзя представить линейным уравнением (в рассматриваемой области), то в многомерном случае, несомненно, будет существенная кривизна. Обратное утверждение, к сожалению, не очевидно.

Итак, **выбор экспериментальной области факторного пространства связан с тщательным анализом априорной информации.**

Поясним наши рассуждения примером .

Пример 1. Изучалось ионообменное разделение смесей группы редкоземельных элементов растворами иминодиуксусной кислоты. Параметр оптимизации — содержание неодима в выходном растворе (элюате) в процентах. Рассматривалось всего два фактора: концентрация элюанта (входного раствора), % вес (\tilde{x}_1) и pH элюанта (\tilde{x}_2). Как построить область определения факторов? Начнем с \tilde{x}_1 . Известно, что при $\tilde{x}_1 > 3$ работать нельзя, так как это предел растворимости данного вещества при нормальной температуре. Значит, верхний предел $\tilde{x}_1 = 3$. С нижним пределом дело обстоит сложнее. Здесь нельзя указать четкую границу. Известно только, что чем ниже концентрация, тем дольше идет процесс. При $\tilde{x}_1 = 0,5$ время протекания процесса находится в разумных пределах. Это и определяет нижнюю границу. Ради большой выгоды ее можно будет сдвинуть, тогда как изменить верхнюю границу практически нельзя.

Для выбора области определения \tilde{x}_2 использовались теоретические представления о процессе, из которых следует, что разделение происходит благодаря одновременному присутствию в системе двух соединений: моно и ди-комплексов. Специальные предварительные опыты показали, что при $\text{pH} < 3$ кислота находится в недиссоциированном состоянии, а при $\text{pH} > 8$ оба соединения разрушаются. Следовательно, \tilde{x}_2 может изменяться от 3 до 8. Если факторы совместимы (а в данном случае это так), то их совместная область определения тоже задана.

Из приведенного примера видно, как непросто решается этот важный вопрос. Теперь в области определения надо найти локальную подобласть для планирования эксперимента. Процедура выбора этой подобласти включает два этапа: выбор основного уровня и выбор интервалов варьирования.

Выбор основного уровня. Наилучшим условиям, определенным из анализа априорной информации, соответствует комбинация (или несколько комбинаций) уровней факторов. *Каждая комбинация является многомерной точкой в факторном пространстве.* Ее можно рассматривать как **исходную точку для построения плана эксперимента.** Назовем ее **основным (нулевым) уровнем.** **Построение плана эксперимента сводится к выбору экспериментальных точек, симметричных относительно нулевого уровня.**

В разных случаях мы располагаем различными сведениями об области наилучших условий. Если имеются сведения о координатах одной наилучшей точки и нет информации о границах определения факторов, то остается рассматривать эту точку в качестве основного уровня. Аналогичное решение принимается, если границы известны и наилучшие условия лежат внутри области.

Положение усложняется, если эта точка лежит на границе (или весьма близко к границе) области. Тогда приходится основной уровень выбирать с некоторым сдвигом от наилучших условий.

Может случиться, что координаты наилучшей точки неизвестны, но есть сведения о некоторой подобласти, в которой процесс идет достаточно хорошо. Тогда основной уровень выбирается либо в центре, либо в случайной точке этой подобласти. Сведения о подобласти можно получить, анализируя изученные ранее подобные процессы, из теоретических соображений или из предыдущего эксперимента.

Наконец, возможен случай с несколькими эквивалентными точками, координаты которых различны. Когда отсутствуют дополнительные данные (технологического, экономического характера и т. д.), выбор произволен. Конечно, если эксперимент недорог и требует немного времени, можно приступить к построению планов экспериментов вокруг нескольких точек.

Следующий пример иллюстрирует одну из возможных ситуаций.

Пример 2. На рис. 4.1 изображена область определения для двух факторов.

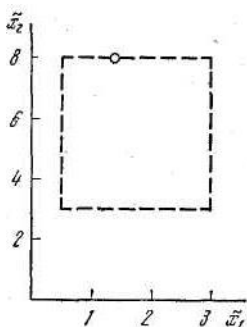


Рис. 4.1 Область определения двух факторов.

Кружком отмечены наилучшие условия, известные из априорной информации. Известно также, что имеется возможность дальнейшего улучшения параметра оптимизации, а данное значение нас не удовлетворяет. Эту точку нельзя рассматривать в качестве основного уровня. Дело в том, что она расположена на границе области определения. Требование симметрии экспериментальных точек относительно нулевого уровня привело бы в этом случае к выходу за границы области определения, чего делать также нельзя. Схема принятия решений при выборе основного уровня приведена на рис. 4.2.

После того как нулевой уровень выбран, переходим к следующему шагу — выбору интервалов варьирования.

Выбор интервалов варьирования. Теперь наша цель состоит в том, чтобы для каждого фактора выбрать два уровня, на которых он будет варьироваться в эксперименте.

Представьте себе координатную ось, на которой откладываются значения данного фактора, для определенности — температуры. Пусть основной уровень уже выбран и равен 100°C . Это значение изображается точкой. Тогда *два интересующих нас уровня можно изобразить двумя точками, симметричными относительно первой*. Будем называть один из этих уровней *верхним*, а второй — *нижним*. Обычно за верхний уровень принимается тот, который соответствует большему значению фактора, хотя это не обязательно, а для качественных факторов вообще безразлично. *Интервалом варьирования факторов называется некоторое число (свое для каждого фактора), прибавление которого к основному уровню дает верхний, а вычитание — нижний уровни фактора*. Другими словами, *интервал варьирования — это расстояние на координатной оси между основным и верхним (или нижним) уровнем*. Таким образом, задача выбора уровней сводится к более простой задаче выбора интервала варьирования.

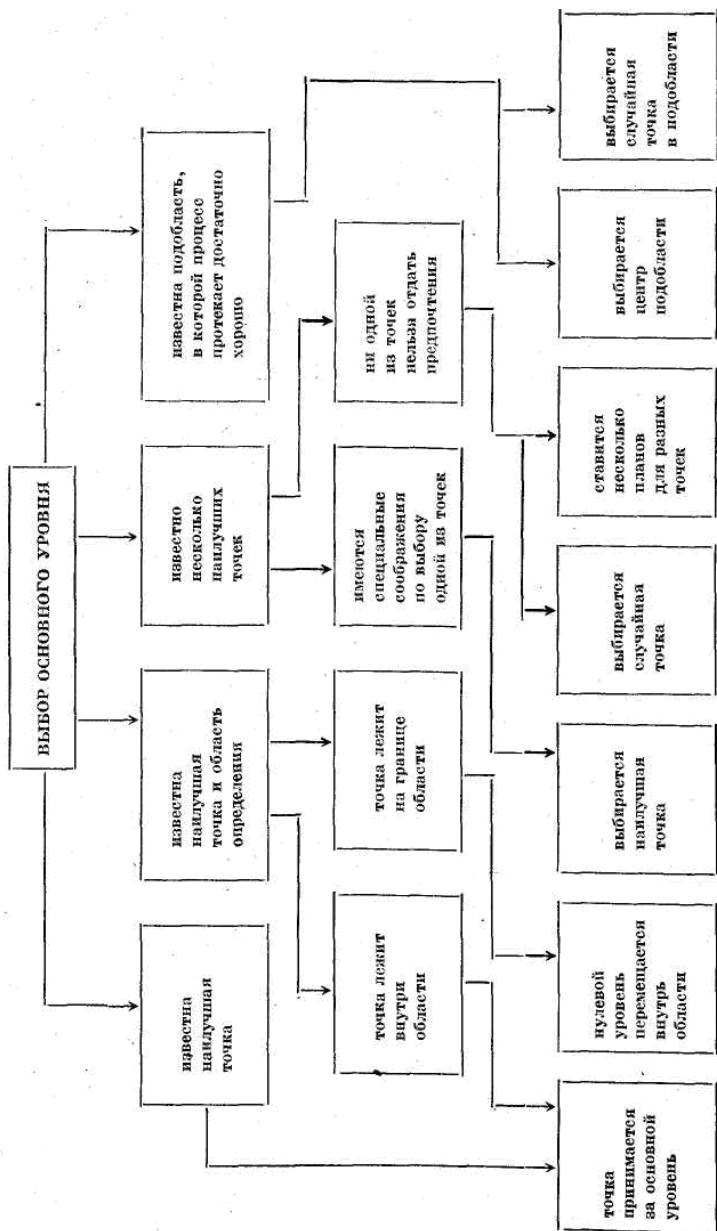


Рис. 4.2. Схема принятия решений при выборе основного уровня

Заметим еще, что для упрощения записи условий эксперимента и обработки экспериментальных данных масштабы по осям выбираются так, чтобы верхний уровень соответствовал +1, нижний —1, а основной — нулю. Для факторов с непрерывной областью определения это всегда можно сделать с помощью преобразования

$$x_j = \frac{\tilde{x}_j - \tilde{x}_{j0}}{I_j},$$

где x_j — кодированное значение фактора, \tilde{x}_j — натуральное значение фактора, \tilde{x}_{j0} — натуральное значение основного уровня, I_j — интервал варьирования, j — номер фактора.

Для качественных факторов, имеющих два уровня, один уровень обозначается +1, а другой — 1; порядок уровней не имеет значения.

Пусть процесс определяется четырьмя факторами. Основной уровень и интервалы варьирования выбраны следующим образом.

	\tilde{x}_1	\tilde{x}_2	\tilde{x}_3	\tilde{x}_4
Основной уровень	3	30	1,5	15
Интервал варьирования	2	10	1	10

Остановимся на первом факторе. Отметим на координатной оси три уровня: нижний, основной и верхний.

Нужно найти кодированное значение для $\tilde{x}_1 = 2,0$. Это значение лежит между 1,0 и 3,0, т. е. между — 1 и 0 в **кодированном** масштабе. Так как в натуральном масштабе 2,0 лежит посередине между 1,0 и 3,0, то ему соответствует — 0,5 в кодированном масштабе. (Для $\tilde{x}_1 = 2,5$ будет $x_1 = -0,25$, для $\tilde{x}_1 = 1,5$ будет $x_1 = -0,75$ и т. д.)

На выбор интервалов варьирования накладываются естественные ограничения сверху и снизу. Интервал варьирования не может быть меньше той ошибки, с которой экспериментатор фиксирует уровень фактора. Иначе верхний и нижний уровни окажутся неразличимыми. С другой стороны, интервал не может быть настолько большим, чтобы верхний или нижний уровни оказались за пределами области определения. Внутри этих ограничений обычно еще остается значительная неопределенность выбора, которая устраняется с помощью интуитивных решений.

Необходимо обратить внимание, что при решении задачи оптимизации мы стремимся выбрать для первой серии экспериментов такую подобласть, которая давала бы возможность для шагового движения к оптимуму. В задачах же интерполяции интервал варьирования охватывает всю описываемую область.

Выбор интервалов варьирования — задача трудная, так как она связана с неформализованным этапом планирования эксперимента. Возникает вопрос, какая априорная информация может быть полезна на данном этапе? Это — сведения *о точности*, с которой экспериментатор фиксирует значения факторов, *о кривизне поверхности отклика* и *о диапазоне изменения параметра оптимизации*. Обычно эта информация является ориентировочной (в некоторых случаях она может оказаться просто ошибочной), но это единственная разумная основа, на которой можно начинать планировать эксперимент. В ходе эксперимента ее часто приходится корректировать.

Точность фиксирования факторов определяется точностью приборов и стабильностью уровня в ходе опыта. Для упрощения схемы принятия решений мы введем приближенную классификацию, полагая, что есть низкая, средняя и высокая точности. Можно, например, считать, что поддержание температуры в реакторе с погрешностью не более 1 % соответствует высокой, не более 5% — средней, а более 10% — низкой точности.

Источником сведений о кривизне поверхности отклика могут служить уже упоминавшиеся графики однофакторных зависимостей, а также теоретические соображения. Из графиков сведения о кривизне можно получить визуально. Некоторое представление о кривизне дает анализ табличных данных, так как *наличию кривизны соответствует непропорциональное изменение параметра оптимизации при равномерном изменении фактора*. Мы будем различать три случая: ***функция отклика линейна, функция отклика существенно нелинейна и информация о кривизне отсутствует***.

Наконец, полезно знать, в *каких диапазонах меняются значения параметра оптимизации в разных точках факторного пространства*. Если имеются результаты некоторого множества опытов, то всегда можно найти наибольшее или наименьшее значения параметра оптимизации. Разность между этими значениями будем называть *диапазоном изменения параметра оптимизации* для данного множества опытов. Условимся различать ***широкий и узкий диапазоны***. Диапазон будет узким, если он существенно отличается от разброса значений параметра оптимизации в повторных опытах. (Этот разброс определяет ошибку опыта.) В противном случае будем считать диапазон широким. Учтем также случай, когда информация отсутствует.

Итак, ***для принятия решений используется априорная информация о точности фиксирования факторов, кривизне***

поверхности отклика и диапазоне изменения параметра оптимизации.

Каждое сочетание градаций перечисленных признаков определяет ситуацию, в которой нужно принимать решение.

При принятых нами градациях возможно $3^3 = 27$ различных ситуаций. Они представлены на рис. 4.3—4.5 в виде кружочков, цифры в которых соответствуют порядковым номерам ситуаций.

Теперь мы приблизились к принятию решения о выборе ***интервалов варьирования***. Для интервалов также введем градацию. Будем рассматривать широкий, средний и узкий интервалы варьирования, а также случай, когда трудно принять однозначное решение. Размер интервала варьирования составляет некоторую долю от области определения фактора. Можно, например, условиться о следующем: если интервал составляем не более 10% от области определения, считать его узким, не более 30% — средним и в остальных случаях — широким. Это, конечно, весьма условно, и в каждой конкретной задаче приходится специально определять эти понятия, которые зависят не только от размера области определения, но и от характера поверхности отклика, и от точности фиксирования факторов.

Перейдем к рассмотрению схем принятия решений. На первой схеме (рис. 4.3) представлены девять ситуаций, имеющих место при низкой точности фиксирования факторов.

При выборе решений учитываются информация о кривизне поверхности отклика и о диапазоне изменения параметра оптимизации. Типичное решение — широкий интервал варьирования. Узкий интервал варьирования совершенно не используется, что вполне понятно при низкой точности.

Пусть ситуация определяется следующими признаками: поверхность отклика линейна, а диапазон изменения параметра оптимизации узок. Какое решение лучше? Эта ситуация обозначена на схеме номером 2. Признаки ситуации определяются стрелками, направленными к данному кружку. Стрелка, выходящая из кружка, указывает решение. Низкая точность фиксирования факторов приводит к отходу от выбора узкого интервала варьирования, иначе результаты могут оказаться неразличимыми. Нам известно, что поверхность линейна. Это не налагает ограничений на расширение интервалов. Кроме того, надо учитывать сведения о диапазоне изменения параметра оптимизации. Он узок, а мы стремимся получить в эксперименте различающиеся значения параметра оптимизации. Поэтому интервал следует увеличивать.

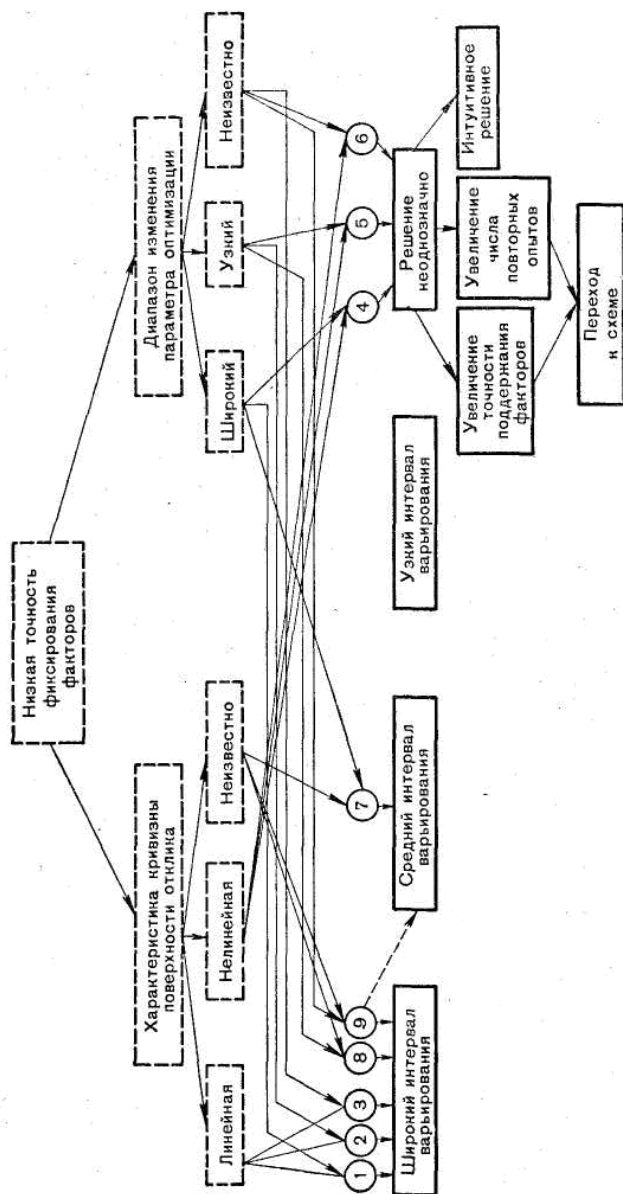


Рис. 4.3. Схема принятия решений при низкой точности фиксирования факторов.

Вернемся снова к схеме. Нам видно, что средний интервал варьирования в этой схеме выбирается дважды, причем в девятой ситуации как редко применяемая альтернатива. Здесь отсутствует информация об обоих признаках, и выбор широкого интервала представляется более естественным.

Наибольшие трудности возникают, когда поверхность отклика нелинейна. Появляется противоречие между низкой точностью фиксирования факторов и кривизной. Первая требует расширения интервала, а вторая — сужения. Решение оказывается неоднозначным. Как поступить в этом случае? Приходится рассматривать дополнительные рекомендации (см. схему). Прежде всего нужно выяснить, нельзя ли увеличить точность эксперимента либо за счет инженерных решений, либо за счет увеличения числа повторных опытов. Если это возможно, то решения принимаются на основе схемы (рис. 4.4) для средней точности фиксирования факторов. Если это невозможно, то для принятия решения нет достаточных оснований и оно становится интуитивным.

Эта схема, как и последующие, служит, весьма грубым приближением к действительности. На практике учитывается еще масса обстоятельств. Например, решения, принимаемые по каждому фактору в отдельности, корректируются при рассмотрении совокупности факторов.

На рис. 4.4 изображена схема для случая средней точности фиксирования факторов.

Характерен выбор среднего интервала варьирования. Лишь в случае нелинейной поверхности и широкого диапазона рекомендуется узкий интервал варьирования. При сочетаниях линейной поверхности с узким диапазоном и отсутствием информации о диапазоне выбирается широкий интервал варьирования. Пунктиром, как и выше, показаны редко применяемые альтернативы.

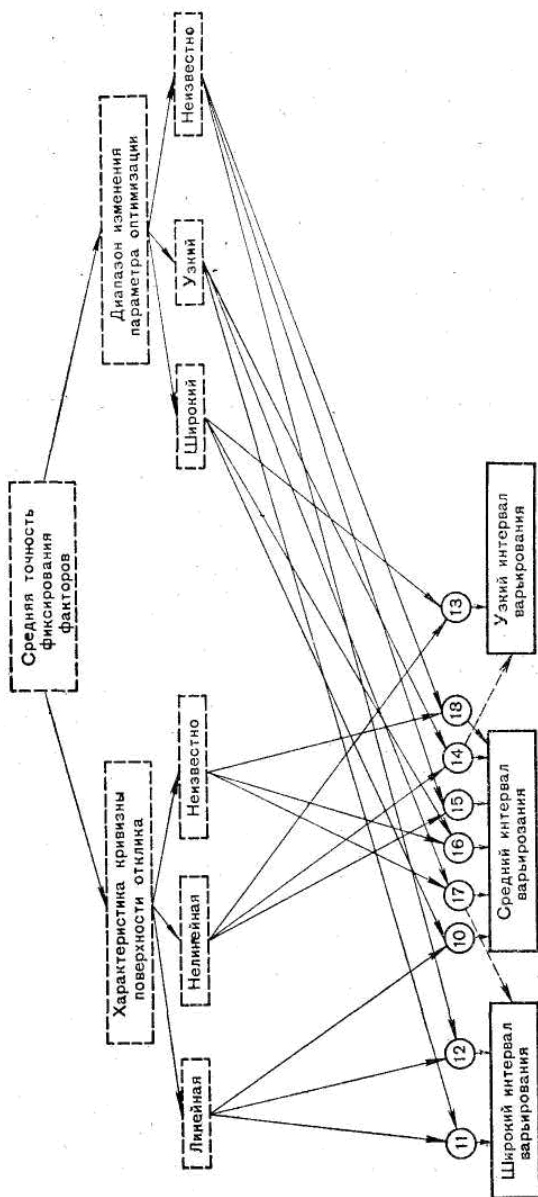


Рис. 4.4. Схема принятия решений при низкой точности фиксирования факторов

Наконец, на рис. 4.5 построена схема для случая высокой точности фиксирования фактора. Сочетание высокой точности с нелинейностью поверхности всегда приводит к выбору узкого интервала. Довольно часто выбирается средний интервал и лишь в двух случаях широкий. В обеих последних схемах отсутствуют неоднозначные решения.

Пример 3. Продолжим рассмотрение примера 1. Область определения факторов была выбрана следующим образом: для \tilde{x}_1 от 0,5 до 3, для \tilde{x}_2 от 3 до 8. Основной уровень: $\tilde{x}_1=1,5$, $\tilde{x}_2=7,0$.

Экспериментатор имел такую априорную информацию: точность фиксирования факторов средняя, поверхность отклика линейна, диапазон изменения параметра оптимизации довольно узок. Этот случай соответствует ситуации 11 схемы (рис. 4.4). Принимаемое решение — широкий интервал варьирования. Экспериментатор выбрал такие интервалы: $I_1=0,5$, $I_2=1,0$, что составляет 20% от области определения факторов. Это несомненно широкие интервалы варьирования. Отметим еще, что для \tilde{x}_2 основной уровень выбран вблизи границы области определения. Поэтому рекомендация о выборе широкого интервала варьирования приводит к совпадению верхнего уровня с этой границей. Так на практике осуществляется выбор интервалов варьирования. Ниже будет продолжено рассмотрение этого примера и мы посмотрим, как оправдываются принятые решения.

Итак, обладая умением выбирать основной уровень и интервалы варьирования факторов, мы можем приступить к построению плана проведения эксперимента.

Но прежде чем приступить к построению плана проведения эксперимента, рассмотрим методы корреляционного, дисперсионного, регрессионного анализов.

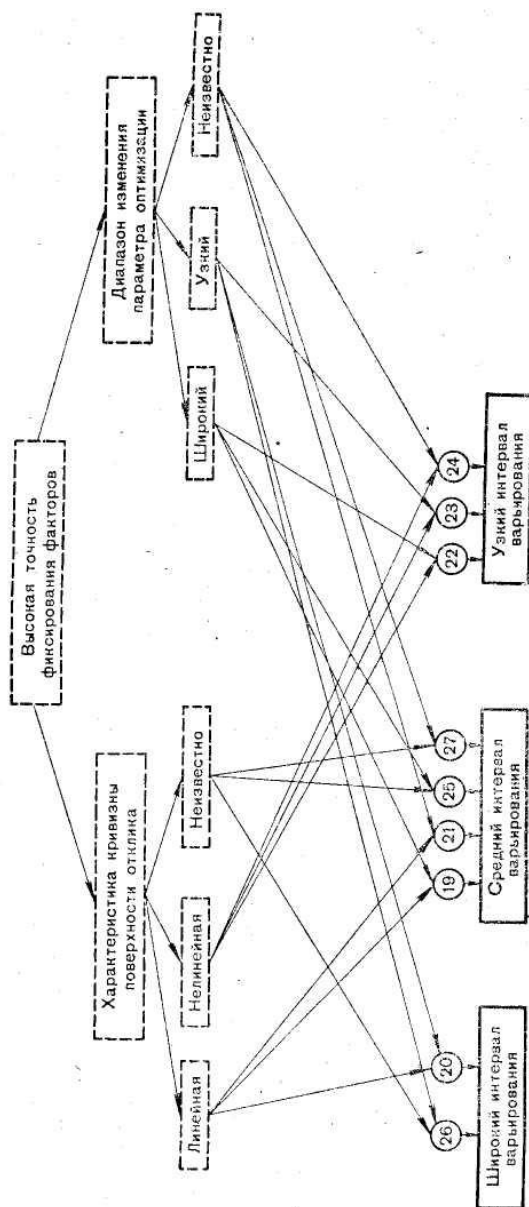


Рис. 4.5. Схема принятия решений при низкой точности фиксирования факторов

4.3. Корреляционный анализ

4.3.1. Понятие о корреляции

При определении общих закономерностей рассматривают, раскрывают функциональные связи, взаимодействия, зависимости; из всех этих связей и зависимостей выделяют главные, наиболее существенные. Раскрытие физической картины связей и взаимодействий помогает выделить наиболее существенные функциональные связи.

Есть более сложные случаи, когда явной функциональной связи между явлениями установить не удастся вследствие или слишком большого разброса результатом опытов, или влияния факторов, физическая роль которых в развитии явления неясна. В этом случае исследователь использует **корреляцию**— *математический прием, позволяющий дать числовую оценку степени взаимной связи между явлениями*. Корреляция иногда также показывает, что между явлениями, возможно, существует еще нескрытая физическая причинная взаимосвязь.

Необходимо помнить, что корреляция является чисто математическим приемом, отнюдь не вскрывающим физики явления; поэтому данная этим приемом числовая оценка связей и зависимостей иногда оказывается формальной, показывающей лишь поверхность явления, а нетипичные опыты дают нетипичную корреляцию.

Как отмечалось, решение задач оптимизации объекта исследования предполагает прежде всего установление количественной связи между переменными, характерными для рассматриваемого объекта.

Ранее был рассмотрен физический подход к нахождению этой связи. Здесь мы проанализируем статистические методы, используемые для этой цели.

Общим для всех статистических методов является то, что исследуемые переменные рассматриваются как случайные величины.

Поэтому такие характеристики случайных величин, как закон распределения, математическое ожидание, дисперсия и т. п., играют в любом из этих методов важную, а иногда и определяющую роль.

Применение статистических методов для определения количественной связи между исследуемыми переменными предполагает наличие двух этапов:

- 1) эксперимент;

2) анализ статистического материала, накопленного в результате эксперимента.

Между этими этапами существует глубокая диалектическая взаимосвязь, условия и способы проведения эксперимента (план эксперимента) предопределяют и дальнейший выбор конкретного метода анализа результатов и, наоборот, выбор определенного метода анализа диктует использование определенного эксперимента.

Точное определение характеристик случайных величин, в частности, количественной связи между различными величинами, возможно только при бесконечном числе наблюдений. Поскольку реальный эксперимент всегда ограничен по своему «объему», анализ его результатов позволяет лишь приближенно оценить исследуемые величины и степень их взаимосвязи. Существующие статистические методы обязательно учитывают это обстоятельство, причем степень достоверности получаемых оценок всегда оговаривается. Собственно, основная задача математической статистики и заключается в том, чтобы на основании ограниченного по объему экспериментального материала, называемого **выборкой**, получать суждения (количественные и качественные) об исследуемых случайных величинах и оценивать степень их достоверности.

Различают две группы математико-статистических методов: статистическая оценка параметров и статистическая проверка гипотез.

Статистическая оценка параметров предусматривает получение оценок (точечных или интервальных) неизвестных параметров генеральной совокупности по параметрам выборки. В результате использования методов этой группы получают суждения типа «коэффициент корреляции между случайными величинами X и Y равен $0,72$ » (пример точечной оценки неизвестного параметра) или «математическое ожидание случайной величины Z с доверительной вероятностью $0,95$ заключено в пределах $27,6 \pm 1,2$ » (пример интервальной оценки неизвестного параметра).

Статистическая проверка гипотез предполагает выдвижение определенных допущений (гипотез) относительно неизвестных качественных (например, тип распределения) или количественных (например, дисперсия) характеристик случайных величин. Правильность этих гипотез проверяется по числовым значениям, полученным из выборки, и в зависимости от результата проверки гипотезы принимают или отвергают. В результате использования методов этой группы получают суждения типа «гипотеза о нормальном распределении случайной величины X не противоречит

экспериментальным данным и, следовательно, может быть принята в качестве рабочей» или «гипотеза о том, что коэффициент корреляции между случайными величинами X и Y равен нулю, с доверительной вероятностью 0,95 не согласуется с экспериментальными данными и должна быть отвергнута», т. е. между X и Y существует линейная зависимость.

Статистические методы с точки зрения их приложения могут быть разделены на методы, направленные на расчет неизвестных характеристик исследуемых случайных величин, и методы, направленные на определение неизвестных взаимосвязей между исследуемыми величинами. Нас будут интересовать последние, так как для управления и оптимизации объекта необходима именно информация о взаимосвязи между характерными величинами. Из методов определения статистической связи между различными величинами наибольшее развитие получили следующие:

- 1) корреляционный;**
- 2) регрессионный;**
- 3) дисперсионный.**

Для того чтобы понять различие между этими методами, проанализируем наиболее простой случай двух величин: X и Y .

В *корреляционном* анализе X и Y — это две равноправные измеримые случайные величины, появляющиеся в виде пар значений при многократном повторении опытов эксперимента. Таким образом, X и Y могут рассматриваться как компоненты двумерной случайной величины (X, Y) , между которыми ищется взаимосвязь. В производстве объектов ситуации такого рода чаще всего возникают при попытке установить связь между структурой и выходными характеристиками изготавливаемых объектов. Например, можно искать связь между парциальным давлением остаточного CO_2 в отпаянном ЭВП и сроком службы прибора. Если такая связь, т. е. корреляция, установлена, то оптимизацию ТП производства данного прибора с точки зрения повышения срока его службы можно проводить, не дожидаясь результатов испытаний на срок службы, а ориентируясь на давление CO_2 в отпаянном приборе. При этом оперативность управления процессом резко повышается.

К *регрессионному* анализу обращаются в том случае, когда есть основания предполагать наличие причинно-следственной связи между измеримыми переменными X и Y . Эксперимент по установлению регрессионной зависимости между переменными внешне может ничем не отличаться от эксперимента по установлению корреляционной зависимости. Однако при регрессионном анализе в отличие от

корреляционного изначально закладывается неравнозначность между X и Y : X считают независимой переменной, а Y — зависимой, принимающей свои значения в зависимости от переменной X . Путем обработки экспериментальных данных (это по-прежнему пары значений, x_i, y_i) отыскиваются неизвестные параметры, устанавливающие количественную связь между значениями зависимой Y и независимой X переменных. Обратимся еще раз к примеру с испытаниями ЭВП на срок службы. Физические предпосылки указывают на наличие причинно-следственной связи между давлением CO_2 в изготовленном приборе и сроком его службы, так как известно, что наиболее частый вид параметрических отказов для ЭВП — это снижение эмиссии катода, а как раз наличие CO_2 в баллоне прибора и вызывает со временем потерю эмиссии за счет отравления катода. Тогда результаты тех же испытаний можно попытаться исследовать с помощью регрессионного анализа. Это приводит к установлению количественной зависимости между сроком службы прибора и парциальным давлением CO_2 в отпаянном приборе. Пользуясь этой зависимостью, можно на последней стадии изготовления ЭВП прогнозировать срок их службы, что дает возможность оперативно оценивать качество ТП и быстро принимать соответствующие меры, когда качество это начинает ухудшаться.

При *дисперсионном* анализе, как и при регрессионном, оценивается степень количественного влияния величины X на значения измеримой величины Y . Отличие заключается в том, что если при регрессионном и корреляционном анализах обе величины являются измеримыми, то при дисперсионном анализе независимая переменная X может быть и не измеримой, хотя и может принимать различные значения. Например, об эффективности автоматизации той или иной технологической операции можно судить по степени уменьшения влияния оператора на выходные параметры процесса. В частности, вопрос может ставиться так: «Зависит ли процент брака на выходе операции от того, кто конкретно проводит данную операцию?» Если такая зависимость сохраняется, то это означает, что уровень автоматизации явно недостаточен. В рассмотренном примере величина X (оператор) не поддается количественному измерению, хотя и может принимать различные значения (Иванов, Петров и т. д.)

Эти же принципиальные отличия между корреляционным, регрессионным и дисперсионным анализами сохраняются и при исследовании связей между многими переменными.

4.3.2. Линейная корреляция

Рассмотрим сначала две случайные величины. Как отмечалось, при корреляционном анализе они предполагаются равнозначными и измеримыми. Задача состоит в обработке результатов наблюдений, в каждом из которых контролируются одновременно значения величин X и Y . Полученные результаты измерений объема n состоят из n пар значений $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, которые можно истолковать как реализации двумерного случайного вектора (X, Y) .

На первом этапе вычисляют среднеарифметические \bar{x} и \bar{y} значения, а также эмпирические дисперсии s_x и s_y :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i; \tag{4.5}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i; \tag{4.6}$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2; \tag{4.7}$$

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \tag{4.8}$$

Эти четыре эмпирических параметра характеризуют измеримые случайные величины X и Y по отдельности: \bar{x} и s_x^2 являются выборочными точечными оценками для математического ожидания μ_x и дисперсии σ случайной величины X ; \bar{y} и s_y^2 — такие же оценки для μ_y и σ_y^2 . При увеличении объема выборки n эти оценки становятся точнее, приближаясь к истинным параметрам μ и σ^2 исследуемой случайной величины:

$$\bar{x} \xrightarrow{n \rightarrow \infty} \mu_x; \quad \bar{y} \xrightarrow{n \rightarrow \infty} \mu_y; \quad s_{x,y}^2 \xrightarrow{n \rightarrow \infty} \sigma_{x,y}^2.$$

Если с увеличением значений x растут (или, наоборот, уменьшаются) значения y , вычислением коэффициента корреляции оценивают, в какой степени связи этих величин приближаются к линейному закону. Корреляция является прямой, если с ростом значения x растут значения y , и обратной, если с ростом значения x значения y уменьшаются. Коэффициент корреляции r показывает тесноту (силу) прямолинейной связи; при установлении его мы ищем наиболее вероятное численное выражение связей.

Общая формула для вычисления коэффициента прямолинейной корреляции между рядами измерений x и y имеет вид:

$$r_{xy} = \frac{\overline{\Delta x \cdot \Delta y}}{\sigma_{\Delta x} \cdot \sigma_{\Delta y}} \quad (4.8-1)$$

При этом

$$\Delta x = x_i - \bar{X} \text{ и } \Delta y = y_i - \bar{Y},$$

т.е. Δx равно разности между данным значением измерения величины x и ее средним значением; аналогично определяют и величину Δy .

Коэффициент корреляции отвечает на вопрос: на сколько сигм в среднем изменится величина y , если величина x изменяется на одну сигму (σ)? Формула (4.8-1) читается так: коэффициент корреляции между рядами значений x и y равен отношению средней произведений Δx на Δy к произведению стандартов величин Δx и Δy .

Оценка прямолинейной корреляции:

- если $r_{xy} = +1,0$ — полная прямая корреляция;
- если $r_{xy} = -1,0$ — полная обратная корреляция;
- чем меньше абсолютное значение r_{xy} , тем меньше сила связи, при $r_{xy} = 0$ связь полностью отсутствует.

Считается, что при $r_{xy} > 0,5$ (по абсолютной величине) корреляционная зависимость достаточно велика и можно говорить о закономерной связи явлений, если это не противоречит сущности явления.

Для оценки надежности полученных значений r_{xy} вычисляют ошибку m_r коэффициента корреляции. Для этого значение r_{xy} возводят в квадрат, полученное вычитают из единицы и результат делят на корень квадратный из количества опытов:

$$m_r = \pm \frac{1 - r_{xy}^2}{\sqrt{n}} \quad (4.8-2)$$

Считается, что значение r_{xy} вполне надежно, если

$$r_{xy} \geq 3m_r.$$

Рассмотрим пример, для которого были ранее найдены некоторые функциональные зависимости. Посмотрим, какова сила связи величинами x и y . Для этого воспользуемся сглаженными данными о среднеквадратических отклонениях величины y и данными величины x (табл. 4.1-1).

Таблица 4.1-1

Общая схема расчетов коэффициента корреляции

x	y	Δx	Δy	$\Delta x \Delta y$	$(\Delta x)^2$	$(\Delta y)^2$
0	1,00	-3,5	-1,01	3,535	12,25	1,011
1	1,60	-2,5	-0,41	1,025	6,25	0,168
2	1,97	-1,5	-0,04	0,060	2,25	0,002
3	2,01	-0,5	0,00	0,000	0,25	0,000
4	2,12	+0,5	+0,11	0,055	0,25	0,012
5	2,21	+1,5	+0,20	0,300	2,25	0,040
6	2,50	+2,5	+0,49	1,225	6,25	0,240
7	2,70	+3,5	+0,69	2,415	12,25	0,475
28	16,11	0	~0	8,615	42,0	1,948

В нижней строке этой таблицы помещены суммарные результаты по столбцам. По этим результатам проведем заключительные вычисления:

$$\bar{x} = \frac{28}{8} = 3,5; \quad \bar{y} = \frac{16,11}{8} = 2,01 \text{ см};$$

$$\overline{\Delta x \Delta y} = \frac{8,615}{8} = 1,077;$$

$$\sigma_{\Delta x} = \sqrt{\frac{42}{8-1}} = 2,447; \quad \sigma_{\Delta y} = \sqrt{\frac{1,948}{8-1}} = 0,527 \text{ см.}$$

Наконец,

$$r_{xy} = \frac{\overline{\Delta x \Delta y}}{\sigma_{\Delta x} \sigma_{\Delta y}} = \frac{1,077}{2,447 \cdot 0,527} = 0,834.$$

Ошибка найденного коэффициента корреляции сравнительно невелика, что указывает на надежность числовой оценки:

$$m_r = \pm \frac{1-0,834^2}{\sqrt{8}} = \pm 0,109, \text{ или около } 11\%.$$

По величине r_{xy} можно заключить, что линейная корреляция между величинами x и соответствующими среднеквадратическими отклонениями величин y достаточно высока и может быть представлена правилом: чем больше x , больше y . В приведенном примере каждому значению x соответствовало одно значение y . На практике при обработке данных, разделенных на классы, каждому классу может соответствовать несколько значений y . В таком случае при расчетах по форме таблицы 4.1-1 проще в столбцах x и y повторять соответствующее количество раз одинаковые данные. Если

количество таких повторений слишком велико, расчеты лучше вести с использованием корреляционных и вспомогательных таблиц.

Рассмотрим случай применения корреляционных таблиц. Чтобы вычислить тесноту связи при многих значениях переменных величин, удобно пользоваться иной формулой значения коэффициента корреляции:

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}. \quad (4.8-3)$$

Здесь σ_x и σ_y — стандарты колеблющихся результатов измерений величин x и y .

Числитель правой части уравнения (4.8-3) носит название ковариации x и y ; его можно найти по удобной для последующих вычислений формуле:

$$C_{xy} = \frac{\sum n_{xy} xy}{n} - \bar{x} \cdot \bar{y}.$$

Здесь n_{xy} — частота некоторой пары значений переменных x и y ;

$\sum n_{xy} xy$ — условное обозначение суммы произведений отдельных сумм по столбцам x (или по строкам y) на соответствующие значения x (или y) в корреляционной таблице;

\bar{x} ; \bar{y} — средние значения переменных x и y ;

n — общее количество наблюдений.

Математический смысл ковариации раскрывается в приводимом ниже примере.

На рисунке 4.5-1 представлены результаты испытаний тракторных двигателей с различной степенью износа гильз, поршней и поршневых колец. При полной подаче топлива и номинальных оборотах измерялись: суммарная относительная неплотность f_s цилиндров дизеля, утечка газов Q из камер сгорания через картер и дымность выхлопа D %. По физическому смыслу за независимую переменную принята величина f_s . Если возьмем точки опытов, то увидим, что каждому среднему значению аргумента f_s может соответствовать не одно определенное значение функций Q и D , а несколько. Такая зависимость будет корреляционной, а сам график называют корреляционным полем. Отметим, что числа масштаба (среднеклассовые значения) стоят против середин клеток чертежа, и поэтому начало отсчетов находится вне чертежа.

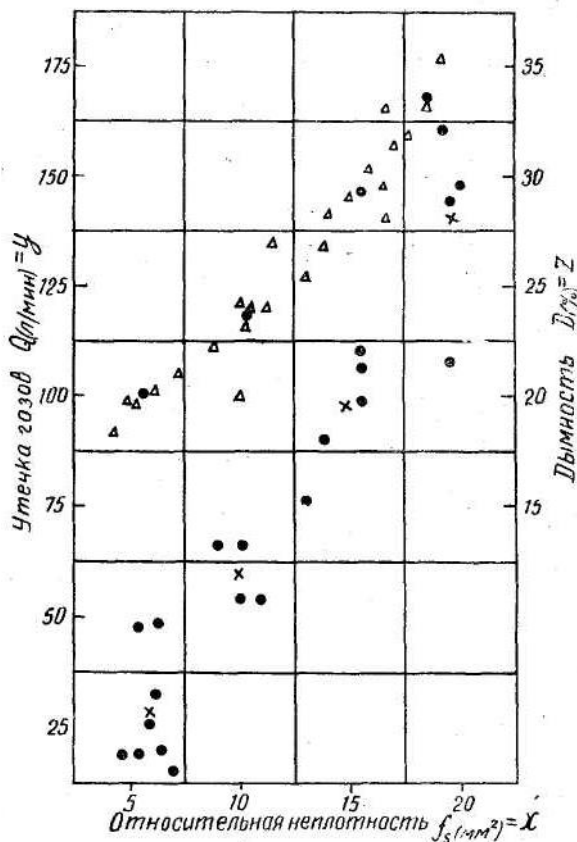


Рис. 4.5.-1. Утечка в картер и дымность газов в выпускной трубе тракторного двигателя в зависимости от суммарной относительной неплотности его цилиндров.

Точками обозначены результаты отдельных измерений величины $Q=y$; крестиками — частные средние этих измерений; треугольниками — результаты измерений величины $D=z$ и кружками — частные средние этих измерений.

Расположение результатов отдельных измерений на рисунке показывает, что налицо связь между величинами f_s , Q и D . Однако не ясно, во-первых, насколько тесна эта связь и, во-вторых, что теснее связано (при данных способах измерений) с замерах неплотностей—

утечка газов или дымление? Ответ может дать сравнение коэффициентов корреляции.

Обозначим величину f_s через x , величины Q и D соответственно через y и z .

В целях лучшего понимания дальнейшего изложения по результатам опытов подготовим данные для проведения сглаженных кривых $y=f_1(x)$; $z=f_2(x)$.

Для удобства их построения определим арифметические частные средние \bar{y}_{x_i} этих зависимостей и нанесем их на чертеж. Например, взяв значения y , расположенные в интервале $x=2,5—7,5$ со средней классовой $x=5$ (восемь измерений), и определив их координаты, получим частную среднюю ординату:

$$\bar{y}_{x_1} = \frac{4,6 \cdot 19 + 5,4 \cdot 19 + 5,4 \cdot 48 + 6,2 \cdot 33,5 + 6,2 \cdot 48,5 + 6,1 \cdot 26 + 6,5 \cdot 20 + 7 \cdot 15}{4,6 + 5,4 + 5,4 + 6,2 + 6,2 + 6,1 + 6,5 + 7} = \frac{1350,8}{47,4} \approx 28,5.$$

Средняя абсцисса этой группы измерений будет

$$\bar{x}_{y_1} = \frac{1350,8}{19 + 19 + 48 + 48,5 + 33,5 + 26 + 20 + 15} \approx 5,9.$$

Далее возьмем группу значений y , расположенного в интервале $x=7,5—12,5$ со средней классовой $x=10$ (четыре измерения), и, вычислив координаты каждого измерения, получим частную среднюю арифметическую ординату:

$$\bar{y}_{x_2} = \frac{8,8 \cdot 66 + 10 \cdot 54 + 10 \cdot 66 + 11 \cdot 54}{8,8 + 10 + 10 + 11} = \frac{2374}{39,8} \approx 59,7.$$

Средняя абсцисса этой группы измерений будет

$$\bar{x}_{y_2} = \frac{2374}{66 + 66 + 54 + 54} = \frac{2374}{240} \approx 9,9.$$

Аналогично находим и остальные координаты: при среднеклассовом значении $\bar{x}=15$, $\bar{y}=97,6$ и $\bar{x}=14,8$: при среднеклассовом значении $x=20$, $\bar{y}_{x_i}=140$ и $\bar{x}_{y_i}=19,6$.

По этим координатам на графике (рис. 4.5-1) построены точки частных средних значений y , отмеченные крестиками.

Таким же образом поступаем и с усреднением данных опытов по величине z . Например, по группе значений z , расположенной в интервале $x=2,5—7,5$ со среднеклассовым значением $x=5$, получаем частную среднюю ординату:

$$\bar{z}_{x_1} = \frac{4,4 \cdot 18,4 + 4,8 \cdot 19,8 + 5,3 \cdot 19,6 + 6,2 \cdot 20,2 + 7,3 \cdot 21}{4,4 + 4,8 + 5,3 + 6,2 + 7,3} = \frac{558}{28} \approx 20.$$

Средняя абсцисса в этом случае будет:

$$\bar{x}_{z_1} = \frac{558}{18,4 + 19,8 + 19,6 + 20,2 + 21} \approx 5,6.$$

Аналогично вычисляем координаты остальных частных средних:

$$\bar{z}_{x_2} = 23,7 \text{ при } \bar{x}_{z_2} = 10,3; \bar{z}_{x_3} = 29,25 \text{ при } \bar{x}_{z_3} = 15,5 \text{ и } \bar{z}_{x_4} = 33,5 \text{ при } \bar{x}_{z_4} = 18,5.$$

Расположение этих средних показано на графике черными точками. Значения частных средних по классам зависят от способа группировки измерений или опытов. Например, если группировку проводить по строкам y , то для $y=12,5$ — $37,5$ со среднеклассовой $y=25$ получим координаты:

$$\bar{y}_{y_1} = \frac{15,7 + 19 \cdot 4,6 + 19 \cdot 5,4 + 20 \cdot 6,5 + 26 \cdot 6,1 + 33,5 \cdot 6,2}{7 + 4,6 + 5,4 + 6,5 + 6,1 + 6,2} = \frac{792}{35,8} \approx 22,1;$$

$$\bar{x}_{y_1} = \frac{792}{15 + 19 + 19 + 20 + 26 + 33,5} \approx 6$$

или при среднеклассовой $y=50$:

$$\bar{y}_{y_2} = \frac{48 \cdot 5,4 + 48 \cdot 5,6,2 + 54 \cdot 10 + 54 \cdot 11}{5,4 + 6,2 + 10 + 11} = \frac{1693}{32,6} \approx 53;$$

$$\bar{x}_{y_2} = \frac{1693}{48 + 48,5 + 54 + 54} \approx 8,3.$$

Это влияние способа группирования на координаты средних, объясняемое разбросом точек измерений, не должно смущать исследователя: хорошо сглаженная кривая, проведенная с учетом полученных средних, при любом способе группирования общих результатов по обоим способам будет давать однозначные значения x и y , x и z или y и z , т.е. каждому значению какой-либо величины будет соответствовать лишь одно значение каждой из других величин (при большом количестве измерений в качестве частных средних берут среднеклассовые значения x и y (или x и z)).

Отметим, что если полученные средние (т. е. крестики или точки) соединить между собой отрезками прямых, то образуются ломаные,

так называемые эмпирические линии регрессии, а сглаженные кривые будут близки к теоретическим линиям регрессии (на рисунке теоретические эмпирические линии регрессии не показаны, чтобы не загромождать рисунок). Для облегчения последующих вычислений составим корреляционные таблицы, в первой строке которых расположим средние по классам значения x , а в первых столбцах — средние по классам значения y , а затем z . Таким образом, вначале определим коэффициент корреляции между x и y , или r_{xy} . В этом примере для наглядности разбивка граф в таблице 4.1-2 и клеток на рисунке 4.5-1 совпадает, так что остается лишь перенести число опытов в клетках в графы таблицы 4.1-2.

Таблица 4.1-2

Корреляционная таблица для x, y

y	x				m_y
	5	10	15	20	
25	6	—	—	—	6
50	2	2	—	—	4
75	—	2	1	—	3
100	—	—	4	1	5
125	—	—	—	—	0
150	—	—	—	3	3
n_x	8	4	5	4	21

Затем суммируют число точек по каждой строке y (m_y) и по каждому столбцу x (n_x).

Далее для вычисления средних значений x и y по корреляционной таблице найдем суммы произведений частот (число измерений в графах таблицы) на соответствующие значения x и y . Разделив общие суммы на общее количество измерений, найдем средние.

Вычисления начнем со столбцов x для определения \bar{y} . В результате получим:

$$\begin{aligned}
 & 6 \cdot 25 + 2 \cdot 50 = 250 \\
 & 2 \cdot 50 + 2 \cdot 75 = 250 \\
 & 1 \cdot 75 + 4 \cdot 100 = 475 \\
 & 1 \cdot 100 + 3 \cdot 150 = 550 \\
 \hline
 & \Sigma y m_y = 1525 \\
 & \bar{y} = \frac{\Sigma y m_y}{n} = \frac{1525}{21} \approx 72,6.
 \end{aligned}$$

Таким же образом находим \bar{x} по строкам y :

$$\begin{aligned} 6 \cdot 5 &= 30 \\ 2 \cdot 5 + 2 \cdot 10 &= 30 \\ 2 \cdot 10 + 1 \cdot 5 &= 35 \\ 4 \cdot 15 + 1 \cdot 20 &= 80 \\ 0 &= 0 \\ 3 \cdot 20 &= 60 \end{aligned}$$

$$\hline \Sigma xn_x = 235$$

$$\bar{x} = \frac{235}{21} \approx 11,2.$$

Затем вычислим по формуле (4.8-4) величину ковариации c_{xy} .

Для получения $\Sigma n_{xy}xy$ перемножим отдельные суммы по столбцам x на соответствующие значения x ; то же сделаем по строкам y . Результаты должны быть одинаковыми:

$$\begin{array}{rcl} 250 \cdot 5 & = & 1\,250 \\ 250 \cdot 10 & = & 2\,500 \\ 475 \cdot 15 & = & 7\,125 \\ 550 \cdot 20 & = & 11\,000 \\ \hline \Sigma n_x xy & = & 21\,875 \end{array} \quad \begin{array}{rcl} 30 \cdot 25 & = & 750 \\ 30 \cdot 50 & = & 1\,500 \\ 35 \cdot 75 & = & 2\,625 \\ 80 \cdot 100 & = & 8\,000 \\ 0 \cdot 125 & = & 0 \\ 60 \cdot 150 & = & 9\,000 \\ \hline \Sigma n_y xy & = & 21\,875 \end{array}$$

Тогда по формуле (4.8-4) получаем:

$$C_{xy} = \frac{21\,875}{21} - 72,6 \cdot 11,2 = 1041,67 - 813,12 = 228,55.$$

Для вычисления стандартов σ_x и σ_y по формуле (1.36) составим таблицу 4.1-3.

Таблица 4.1-3

Определение данных для вычисления σ_x и σ_y

x_i	n_{x_i}	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_{x_i} \frac{(x_i - \bar{x})^2}{x_i}$	y_i	m_{y_i}	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$m_{y_i} \frac{(y_i - \bar{y})^2}{y_i}$
5	8	-6,2	38,44	307,52	25	6	-47,6	2265,76	13 594,56
10	4	-1,2	1,44	5,76	50	4	-22,6	510,76	2 043,04
15	5	+3,8	14,44	72,20	75	3	+2,4	5,76	17,28
20	4	+8,8	77,44	309,76	100	5	+27,4	750,76	3 753,80
—	—	—	—	—	125	0	+52,4	2745,76	0
—	—	—	—	—	150	3	+77,4	5990,76	17 972,28
Σ	21	—	—	695,24	Σ	21	—	—	37 380,96

Отсюда получаем:

$$\sigma_{x_i} = \sqrt{\frac{695,24}{21-1}} \approx 5,90; \sigma_{y_i} = \sqrt{\frac{37\,380,96}{21-1}} \approx 43,23.$$

Далее находим по формуле (4.8-3) значение коэффициента корреляции

$$r_{xy} = \frac{228,55}{5,9 \cdot 43,23} \approx 0,897$$

и по формуле (4.8-2) его ошибку

$$m_r = \pm \frac{1 - 0,897^2}{\sqrt{21}} \approx \pm 0,044.$$

Эта ошибка мала, и значение r_{xy} вполне надежно, так как

$$0,897 > 3 \cdot 0,044.$$

Этим примером вычисления ошибки коэффициента корреляции мы и ограничимся.

Как видно, прямолинейная связь кривой y (утечки газов) с независимой переменной x (относительная неплотность) весьма тесна. В самом деле, сглаженная кривая (близкая к кривой, носящей название теоретической линии регрессии) может быть выражена уравнением Q , л/мин = $4,4 f^{1,15}$; показатель степени мало отличается от единицы. В соответствии с физическим смыслом явления при $f_s=0$ $Q=0$. Но колебания результатов измерения здесь велики: для величины x коэффициент вариации

$$v_x = \frac{5,90}{11,2} 100 \approx 52,7\%,$$

а для величины y

$$v_y = \frac{43,23}{72,6} 100 = 59,6\%,$$

то есть еще больше. Так как сами по себе примененные для измерений приборы имеют высокую точность, делаем вывод, что утечка газов является сложной функцией, зависящей не только от рассматриваемых здесь факторов.

Составим такую же корреляционную таблицу для $D = z$ и выполним аналогичные вычисления.

Таблица 4.1-4

Корреляционная таблица для x и z

z	x				m_z
	5	10	15	20	
20	5	2	—	—	7
25	—	5	2	—	7
30	—	—	6	1	7
35	—	—	1	2	3
n_x	5	7	9	3	24

Найдем среднюю \bar{z} :

$$\begin{aligned}
 & 5 \cdot 20 = 100 \\
 & 2 \cdot 20 + 5 \cdot 25 = 165 \\
 & 2 \cdot 25 + 6 \cdot 30 + 1 \cdot 35 = 265 \\
 & 1 \cdot 30 + 2 \cdot 35 = 100 \\
 & \hline
 & \Sigma z m_z = 630 \\
 & \bar{z} = \frac{630}{24} \approx 26,25.
 \end{aligned}$$

Определим среднюю \bar{x} :

$$\begin{aligned}
 & 5 \cdot 5 + 2 \cdot 10 = 45 \\
 & 5 \cdot 10 + 2 \cdot 10 = 80 \\
 & 6 \cdot 15 + 1 \cdot 20 = 110 \\
 & 1 \cdot 15 + 2 \cdot 20 = 55 \\
 & \hline
 & \Sigma x n_x = 290 \\
 & \bar{x} = \frac{290}{24} \approx 12,08.
 \end{aligned}$$

Далее вычислим Σn_{xz} :

$$\begin{array}{ll}
 100 \cdot 5 = 500 & 45 \cdot 20 = 900 \\
 165 \cdot 10 = 1650 & 80 \cdot 25 = 2000 \\
 265 \cdot 15 = 3975 & 110 \cdot 30 = 3300 \\
 100 \cdot 20 = 2000 & 55 \cdot 35 = 1925 \\
 \hline
 \Sigma n_x xz = 8125 & \Sigma n_z xz = 8125
 \end{array}$$

По формуле (4.8-4) получаем ковариацию

$$C_{xz} = \frac{8125}{24} - 26,25 \cdot 12,08 = 21,35.$$

Составим таблицу 4.1-5 и найдем стандарты σ'_x и σ'_z .

Таблица 4.1-5

Определение данных для вычислений σ'_x и σ'_z .

x_i	n_{xi}	\bar{x}	$(x_i - \bar{x})^2$	$n_{xi}(x_i - \bar{x})^2$	z_i	m_{zi}	\bar{z}	$(z_i - \bar{z})^2$	$m_{zi}(z_i - \bar{z})^2$
5	5	-7,08	50,13	250,65	20	7	-6,25	39,06	273,42
10	7	-2,08	4,33	30,31	25	7	-1,25	1,56	10,92
15	9	+2,92	8,53	76,77	30	7	+3,75	14,06	98,42
20	3	+7,92	62,73	188,19	35	3	+8,75	76,56	229,68
Σ	24	—	—	545,92	Σ	24	—	—	612,44

Из таблицы 4.1-5

$$\sigma'_x = \sqrt{\frac{545,92}{24-1}} \approx 4,87; \quad \sigma'_z = \sqrt{\frac{612,44}{24-1}} \approx 5,16,$$

поэтому

$$r_{xz} = \frac{21,35}{4,87 \cdot 5,16} \approx 0,852.$$

Сравнивая величины r_{xy} и r_{xz} , приходим к выводу, что кривая $z=D$ менее тесно связана прямолинейной связью с величиной $x=f_s$. Но, может быть, это зависит от разброса точек? Определим коэффициенты вариации;

$$v_x = \frac{4,87}{12,08} 100 = 39,5\%; \quad v_z = \frac{5,16}{26,25} 100 = 19,8\%.$$

Как видно, изменчивость результатов измерений здесь значительно меньшая, чем в первом случае, а разброс результатов измерений ординат незначительный. Таким образом, первый вывод, который следует сделать из изложенного, таков: кривая $z=D$ больше отклоняется от прямой, чем кривая $y=Q$. Действительно, сглаженной кривой с небольшой ошибкой удовлетворяет уравнение

$$D = 0,2 f_s^{1,5} + 16,3.$$

Далее, дымление присуще двигателям и при отсутствии неплотностей (вернее, при весьма малой неплотности), и для данных условий в этих случаях $D \approx 16\%$. Можно заметить, что значения

\bar{x} и σ_x меняются в зависимости от того, рассчитаны ли эти значения для связи с y или для связи с z . Это объясняется различной измерений y и z , а следовательно, и различной их взаимосвязью с величиной x .

В приведенном примере вычисление ошибок связано со среднеклассовыми значениями функций, как наиболее вероятными.

Если вычисления провести по всем отдельным точкам (несглаженным), ошибка будет больше, это означает, что расчет идет не только по наиболее вероятным точкам кривых закономерностей.

Отметим еще, что в литературе встречаются и другие обозначения эмпирической ковариации. Приведем одно из них. Как мы знаем, нас интересует наличие связи между X и Y , а также сила (теснота) этой связи. Информацию об этом дает эмпирическая ковариация между значениями x и y , которая определяется следующим образом:

$$\widehat{\text{cov}}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}). \quad (4.9)$$

Нормируя $\widehat{\text{cov}}(x, y)$ с помощью среднеквадратичных отклонений s_x и s_y , получают эмпирический коэффициент корреляции между значениями x и y :

$$r_{xy} = \frac{\widehat{\text{cov}}(x, y)}{s_x s_y}. \quad (4.10)$$

Так как $\widehat{\text{cov}}(x, y) = \widehat{\text{cov}}(y, x)$, то $r_{xy} = r_{yx}$, т.е. порядок индексов не важен, что и должно иметь место, так как случайные величины X и Y полагаем равноправными.

Эмпирический коэффициент r_{xy} — это выборочная оценка для истинного коэффициента корреляции ρ_{xy} , который является мерой силы (тесноты) и направления линейной связи между значениями компонент случайного вектора (X, Y) .

Если $|\rho_{xy}| = 1$, то существует линейная функциональная зависимость между x и y , т.е. в любой реализации случайного вектора (X, Y) компоненты x и y связаны одним и тем же коэффициентом пропорциональности γ , так что

$$(x_i - \mu_x) = \gamma (y_i - \mu_y), \quad i = 1, 2, \dots$$

В этом случае результаты наблюдений случайного вектора (X, Y) в выборке объема n находятся на одной прямой (рис. 4.5, а, б).

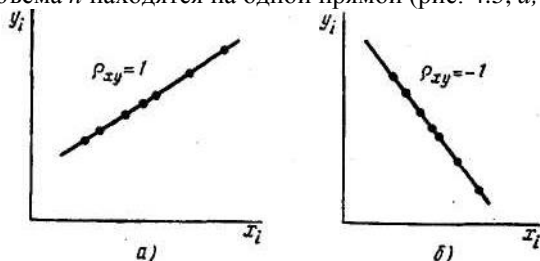


Рис. 4.5.

Если $0 < |\rho_{xy}| < 1$, то существует линейная связь между значениями x и y , внутренняя природа которой более сложна, чем простая линейная функциональная зависимость (это может быть случай, когда $X=X_1 + X_2$, $Y=Y_1+Y_2$ и между X_1 и Y_1 имеется линейная функциональная зависимость, а X_2 и Y_2 — независимые случайные добавки). Результаты наблюдений случайного вектора (X, Y) в выборке объема n группируются вблизи некоторой прямой (рис. 4.6, а, б).

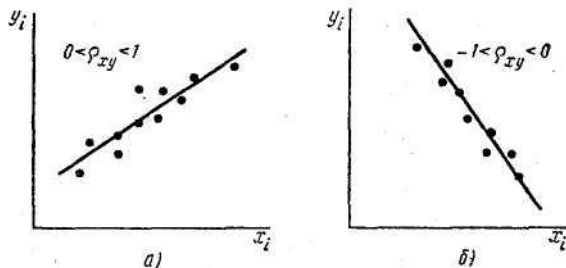


Рис. 4.6.

Если $\rho_{xy}=0$, то невозможно в общем случае сделать вывод о независимости случайных величин X и Y . Можно лишь утверждать, что линейная связь отсутствует (нелинейная связь при этом не исключается, как это видно из рис. 4.7, а, б).

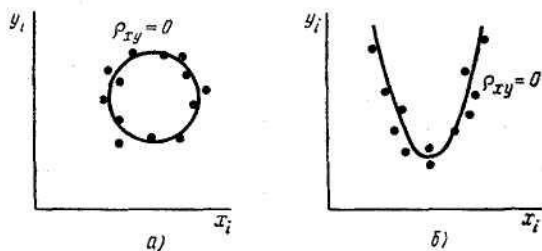


Рис. 4.7.

Однако существуют случаи, когда равенство $\rho_{xy}=0$ означает независимость случайных величин X и Y , например, если случайные величины X и Y имеют нормальное или равномерное распределение.

Из рассмотрения свойств коэффициента корреляции можно сделать вывод о том, что корреляционный анализ наиболее эффективен, когда есть основания предполагать преобладающий характер линейной связи между исследуемыми переменными. Применение корреляционного анализа в случае, когда возможна нелинейная зависимость между переменными, нецелесообразно. Однако использование физических представлений о природе исследуемых величин может расширить

область применения корреляционного анализа за счет предварительного преобразования переменных. Например, анализ физики теплового пробоя диэлектриков приводит к следующему соотношению между пробивным напряжением $U_{пр}$ и толщиной диэлектрика d :

$$U_{пр} \approx AV\sqrt{d}. \quad (4.11)$$

Учитывая это, можно подвергнуть корреляционному анализу следующие величины: а) $X = U_{пр}^2$, $Y = d$; б) $X = U_{пр}$, $Y = \sqrt{d}$. Какой из рассмотренных вариантов преобразования предпочесть? Общей рекомендацией на этот счет является следующая: преобразование исследуемых величин должно быть таким, чтобы закон распределения преобразованных величин был по возможности ближе к закону нормального распределения. При этом облегчается проверка статистической гипотезы о значимости корреляции (о проверке статистических гипотез см. далее).

В приведенном примере следует остановиться на первом варианте, так как обычно линейные размеры изготавливаемых структур подчиняются закону нормального распределения (хотя и не всегда). Если физическая модель теплового пробоя (4.11) верна, то и $U_{пр}^2$ также будет иметь нормальное распределение. Последнее, однако, требует проверки, так как может оказаться, что в проводимых исследованиях пробой имеет не тепловой, а электрический характер, при котором $U_{пр}$ пропорционально d , а следовательно, в качестве исследуемых величин следует взять $X = U_{пр}$; $Y = d$. Данный пример свидетельствует о том, что прежде чем обратиться к статистическим методам анализа, нужно тщательно проанализировать всю априорную информацию и изучить все возможные варианты исследования.

Допустим, что при обработке эксперимента получен эмпирический коэффициент корреляции $r_{xy} = 0,52$. Можно ли считать установленной наличие линейной связи между исследуемыми случайными величинами? Не будем торопиться с выводами. Вспомним, что выборочные оценки сходятся к оцениваемым параметрам генеральной совокупности лишь при неограниченном возрастании объема выборки n . При любом конечном n они могут флуктуировать от выборки к выборке, отличаясь порой значительно от истинного значения оцениваемого параметра. Возможные отличия тем больше, чем меньше объем выборки, поэтому может оказаться, что в нашем случае $r_{xy} = 0$, а полученное значение $r_{xy} = 0,52$ — не что иное, как вполне естественное (статистически незначимое) отклонение от истинного коэффициента корреляции, вызванное малым числом наблюдений.

Для решения вопроса о статистической значимости полученного выборочного коэффициента корреляции выдвигают гипотезу H_0 : $\rho_{xy} = 0$, а затем проверяют, не противоречит ли она экспериментальным данным. Эта проверка легко осуществляется, если известен закон распределения исследуемых случайных величин X и Y . Тогда по известному закону распределения и объему выборки n находят критическое значение эмпирического коэффициента корреляции, вероятность превышения которого в рамках гипотезы H_0 не может быть больше заданного малого числа α , называемого **уровнем значимости**. Если найденный из эксперимента эмпирический коэффициент корреляции r_{xy} превысил это критическое значение, то тем самым осуществилось событие, маловероятное в рамках выдвинутой гипотезы H_0 : $\rho_{xy} = 0$. Это дискредитирует гипотезу H_0 , и ее отвергают, т. е. считают, что r_{xy} значимо отличается от нуля, так что $\rho_{xy} \neq 0$. Следовательно, существует линейная связь между переменными X и Y . Доверительная вероятность такого заключения равна $1 - \alpha$. Если распределение исследуемых величин X и Y можно считать нормальным, то проверку значимости эмпирического коэффициента корреляции осуществляют по табл. 4.1, содержащей рассчитанные заранее наибольшие случайные значения коэффициента корреляции.

Таблица 4.1

Число степеней свободы $n-2$	Уровень значимости α		
	0,05	0,01	0,0027
5	0,75	0,87	0,93
10	0,58	0,71	0,78
15	0,48	0,61	0,68
20	0,42	0,53	0,61
25	0,38	0,49	0,55
30	0,35	0,45	0,51
35	0,32	0,42	0,48
40	0,30	0,39	0,45
50	0,27	0,35	0,41
60	0,25	0,33	0,37
70	0,23	0,30	0,35
80	0,22	0,28	0,33
90	0,21	0,26	0,31
100	0,19	0,25	0,29
150	0,16	0,21	0,24
200	0,14	0,18	0,21
300	0,11	0,15	0,17
500	0,09	0,11	0,13
1000 и более	0,05	0,09	0,1

Проверим, например, значим ли эмпирический коэффициент корреляции $r_{xy} = 0,52$, если он получен из выборки объема $n = 22$. Для $t = n - 2 = 20$ можно вынести следующее суждение: «с доверительной вероятностью 0,95 корреляция значима ($r_{xy} > 0,42$)». Для $t = 20$ нельзя гарантировать значимость корреляции с большей доверительной вероятностью, так как $r_{xy} < 0,53$ и тем более $r_{xy} < 0,61$.

Проверим, какой вывод можно сделать, если значение $r_{xy} = 0,52$ было получено из выборки объема $n = 12$. Можно сделать вывод о том, что значение $r_{xy} = 0,52$ в столь малой по объему выборке не позволяет утверждать с достаточно высокой доверительной вероятностью о наличии корреляции между исследуемыми величинами. При этом не отвергается наличие корреляции, а только утверждается, что даже такая большая цифра, как $r_{xy} = 0,52$, на малой выборке не позволяет определить существование корреляции, так как это значение, возможно, является случайной флуктуацией, вполне допустимой в рамках гипотезы $H_0: \rho_{xy} = 0$. Внимательный анализ табл. 4.1 позволяет сделать вывод о том, что в тех случаях, когда коэффициент корреляции ρ_{xy} мал ($\rho < 0,2$), определить его с достаточной достоверностью можно только на выборках большого объема ($n > 100$).

После того как установлена значимость коэффициента корреляции, он может быть использован для прогнозирования значений одной из случайных величин, если известно значение второй. Это прогнозирование основывается на следующих формулах:

$$M(x/y) = M(x) - (y - M(y)) \sigma_x \frac{\rho_{xy}}{\sigma_y}; \quad (4.12)$$

$$\sigma(x/y) = \sqrt{1 - \rho_{xy}^2} \sigma_x, \quad (4.13)$$

где $M(x/y)$ — условное математическое ожидание величины x при фиксированной величине y ; $\sigma(x/y)$ — условное среднеквадратичное отклонение величины x при фиксированной y .

Из (4.13) следует, что чем ближе коэффициент корреляции к единице, тем точнее могут быть предсказаны значения x по известным значениям y . Именно это обстоятельство позволяет использовать корреляционный анализ для управления объектами.

Проведенный анализ для двух переменных X и Y легко обобщается на случай многих переменных. Рассмотрим многомерный случайный вектор \mathbf{X} , реализациями которого являются наборы значений $(x_{1i}, x_{2i}, \dots, x_{pi})$, $i = 1, 2, \dots, n$. Вычисляя попарные ковариации между компонентами этого случайного вектора $\widehat{\text{cov}}(x_k, x_l)$, образуют эмпирическую матрицу ковариации:

$$\widehat{\mathbf{V}}_X = \begin{bmatrix} s_{x_1}^2 & \widehat{\text{COV}}(x_1, x_2) & \dots & \widehat{\text{COV}}(x_1, x_p) \\ \widehat{\text{COV}}(x_2, x_1) & s_{x_2}^2 & \dots & \widehat{\text{COV}}(x_2, x_p) \\ \dots & \dots & \dots & \dots \\ \widehat{\text{COV}}(x_p, x_1) & \widehat{\text{COV}}(x_p, x_2) & \dots & s_{x_p}^2 \end{bmatrix}. \quad (4.14)$$

Эта матрица симметрична относительно главной диагонали, так как $\widehat{\text{COV}}(x_k, x_l) = \widehat{\text{COV}}(x_l, x_k)$; на главной диагонали находятся эмпирические дисперсии компонентов вектора \mathbf{X} .

Нормируя элементы этой матрицы $\widehat{\text{COV}}(x_k, x_l)$ с помощью среднеквадратичных отклонений s_{x_k} и s_{x_l} , получают эмпирическую корреляционную матрицу

$$\mathbf{R}_X = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}. \quad (4.15)$$

Это симметричная матрица, на главной диагонали которой располагаются единицы.

Важным свойством корреляционной матрицы является то, что ее определитель удовлетворяет неравенству

$$0 \leq |\mathbf{R}_X| \leq 1. \quad (4.16)$$

Определитель корреляционной матрицы характеризует глубину связи между компонентами вектора \mathbf{X} : если $|\mathbf{R}_X| \rightarrow 1$, то компоненты вектора \mathbf{X} независимы, если $|\mathbf{R}_X| \rightarrow 0$, то значение отдельного компонента является линейной комбинацией всех остальных. Последнее возможно даже в случае, когда в матрице \mathbf{R}_X нет больших значений r_{hl} , т. е. корреляция может быть существенной в целом.

Корреляционную матрицу часто используют для сокращения числа выходных переменных, учитываемых при управлении объектом. Для этого сначала разбивают выходные параметры на группы так, чтобы в одну группу попали близкие параметры (сильно коррелированные), а в разные — параметры, слабо связанные между собой. Затем внутри каждой группы выбирают один или несколько параметров, а остальные исключают. Таким образом, управление объектом осуществляют, ориентируясь на оставленные параметры, общее число которых существенно меньше исходного числа параметров, что облегчает

управление. В то же время сильная корреляция между исключенными и оставленными параметрами автоматически обеспечивает нужный эффект управления и для исключенных параметров.

4.3.3. Нелинейная корреляция

Приведем понятие о нелинейной корреляции между двумя величинами, для чего воспользуемся только рассмотренной выше связью z и x как более выраженной криволинейной.

Общей мерой тесноты нелинейной корреляционной связи между двумя величинами является корреляционное отношение z к x :

$$\eta_{zx} = \frac{\sigma(\bar{z}_x)}{\sigma_z} \tag{4.16-1}$$

Эта формула отвечает на вопрос: какая часть полной колеблемости z обусловлена изменчивостью величины x ?

В формуле (4.16-1) σ_z находят так, как было показано выше; числитель же правой части этой формулы (стандарт частных средних \bar{z}_x) представлен так:

$$\sigma(\bar{z}_x) = \sqrt{\frac{\sum n_{xi}(\bar{z}_{xi} - \bar{z})^2}{n - 1}}, \tag{4.16-2}$$

где \bar{z}_{xi} — частные средние значения, z , определяемые для каждого интервала со среднеклассным значением x , что уже было выполнено ранее (см. Нахождение средних ординат). Вычислим по формуле

(4.16-2) значение $\sigma(\bar{z}_x)$, учитывая, что $\bar{z} = 26,25$. Составим таблицу 4.2-1.

Таблица 4.2-1

Данные для вычисления $\sigma\bar{z}_x$

\bar{z}_{xi}	$z_{xi} - \bar{z}$	$(\bar{z}_{xi} - \bar{z})^2$	n_{xi}	$n_{xi}(\bar{z}_{xi} - \bar{z})^2$
20,00	-6,25	39,06	5	195,30
23,70	-2,55	6,50	7	45,50
29,25	+3,00	9,00	9	81,00
33,50	+7,25	52,56	3	157,68
Σ	—	—	24	479,48

Отсюда

$$\sigma(\bar{z}_x) = \sqrt{\frac{479,48}{24-1}} \approx 4,57.$$

Так как ранее было получено значение $\sigma_z=5,16$, то по формуле (4.16-1) получим:

$$\eta_{zx} = \frac{4,57}{5,16} \approx 0,886.$$

Корреляционное отношение всегда положительно и находится между 0 и 1. Если оно равно нулю, корреляционная связь отсутствует. Чем больше значение корреляционного отношения, тем теснее корреляционная связь. Если значение корреляционного отношения совпадает со значением коэффициента корреляции, связь линейна. В других случаях $\eta > r$. Для нелинейных зависимостей корреляционное отношение является единственно правильным измерителем тесноты связи.

4.3.4. Множественная линейная корреляция

До сих пор разбиралась корреляция между двумя величинами: y и x , а также z и x . Но в ряде случаев приходится оценивать взаимосвязь между несколькими величинами, что является предметом множественной корреляции. Наиболее важное практическое значение имеет линейная множественная корреляция.

Мы остановимся на простом случае линейной корреляции между тремя величинами:

$$f_s = x, Q = y \text{ и } D = z.$$

При этом должны быть известны коэффициенты корреляции r_{xy} , r_{xz} и r_{yz} . Тогда мерой тесноты связи будет служить сводный (общий, совокупный) коэффициент множественной корреляции R . Он всегда положителен и заключен между 0 и 1.

Для определения тесноты связи между z и x , y в совокупности величина R вычисляется по уравнению:

$$R_{z,xy} = + \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xy}r_{xz}r_{yz}}{1-r_{xz}^2}}. \quad (4.16-3)$$

Для установления тесноты связи между y и x, z величина R будет равна:

$$R_{y,xz} = + \sqrt{\frac{r_{xy}^2 + r_{yz}^2 - 2r_{xy}r_{xz}r_{yz}}{1-r_{xz}^2}}.$$

Аналогично составляется уравнение для $R_{x,yz}$.

Приведем пример вычисления коэффициента множественной корреляции, взяв за основу данные рисунка 4.5-1 и предыдущих вычислений коэффициентов корреляции.

Прежде всего необходимо найти значение коэффициента корреляции r_{yz} между величинами y и z . При составлении корреляционной таблицы для y, z используем средние, полученные для облегчения проведения сглаженных кривых. Таким образом, выпишем для корреляционной таблицы данные о количестве измерений величины z , находящиеся в столбцах: $\bar{y}_{x1}, \bar{y}_{x2}, \bar{y}_{x3}$ и \bar{y}_{x4} .

Кроме того, отнесем полученное количество измерений величины Z к тем среднеклассовым значениям y , которым соответствуют $\bar{y}_{x1}, \bar{y}_{x2}, \bar{y}_{x3}$ и \bar{y}_{x4} . Тогда корреляционная таблица для y, z примет следующий вид (таб. 4.2-2).

Таблица 4.2-2

Корреляционная таблица для y, z

z	y						m_z
	25	50	75	100	125	150	
20	5	2	—	—	—	—	7
25	—	5	—	2	—	—	7
30	—	—	—	6	—	1	7
35	—	—	—	1	—	2	3
n_y	5	7	0	9	0	3	24

Вычислим \bar{z} :

$$\begin{aligned}
 & 5 \cdot 20 = 100 \\
 & 2 \cdot 20 + 5 \cdot 25 = 165 \\
 & \quad \quad \quad 0 = 0 \\
 & 2 \cdot 25 + 6 \cdot 30 + 1 \cdot 35 = 265 \\
 & \quad \quad \quad 0 = 0 \\
 & 1 \cdot 30 + 2 \cdot 35 = 100 \\
 & \hline
 & \Sigma ym_z = 630 \\
 & \bar{z} = \frac{630}{24} \approx 26,25.
 \end{aligned}$$

Далее определим \bar{y} :

$$\begin{aligned}
 5 \cdot 25 + 2 \cdot 50 &= 225 \\
 5 \cdot 50 + 2 \cdot 100 &= 450 \\
 6 \cdot 100 + 1 \cdot 150 &= 750 \\
 1 \cdot 100 + 2 \cdot 150 &= 400 \\
 \hline
 \Sigma zn_y &= 1825 \\
 \bar{y} &= \frac{1825}{24} \approx 76,00.
 \end{aligned}$$

Вычисление средних здесь проверочное, причем значение y отклоняется незначительно от ранее полученного.

Проводим вычисления для нахождения ковариации C_{yz} :

$$\begin{array}{r}
 100 \cdot 25 = 2\,500 \qquad 225 \cdot 20 = 4\,500 \\
 165 \cdot 50 = 8\,250 \qquad 450 \cdot 25 = 11\,250 \\
 0 \cdot 75 = 0 \qquad 750 \cdot 30 = 22\,500 \\
 265 \cdot 100 = 26\,500 \qquad 400 \cdot 35 = 14\,000 \\
 0 \cdot 125 = 0 \qquad \Sigma n_z yz = 52\,250 \\
 100 \cdot 150 = 15\,000 \\
 \hline
 \Sigma n_y yz = 52\,250 \\
 C_{yz} = \frac{52\,250}{24} - 26,25 \cdot 76,00 \approx 181,92.
 \end{array}$$

Вычислим, как и ранее, значение σ'_y , так как группировка y_i изменилась (по сравнению с ранее произведенными в табл. 4.1-3 вычислениями). Для этого составляем таблицу 4.2-3.

Таблица 4.2-3

Определение данных для вычисления σ'_y

y_i	n_{yi}	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$n_{yi} (y_i - \bar{y})^2$
25	5	-51	2601	13 005
50	7	-26	676	4 732
75	0	-1	1	0
100	9	+24	576	5 184
125	0	+49	2401	0
150	3	+74	5476	16 428
Σ	24	-	-	39 349

Отсюда

$$\sigma'_y = \sqrt{\frac{39\,349}{24 - 1}} \approx 41,4.$$

Группировка z_i по сравнению с ранее полученными данными (см. табл. 4.1-5) не изменилась, и поэтому берем $\sigma_z = 5,16$. Отсюда

$$r_{yz} = \frac{181,92}{41,4 \cdot 5,16} \approx 0,849.$$

Наконец, по формуле (4.16-3) определяем тесноту связи между z и x , y :

$$R_{z,xy} = + \sqrt{\frac{0,852^2 + 0,849^2 - 2 \cdot 0,897 \cdot 0,852 \cdot 0,849}{1 - 0,897^2}} \approx 0,874.$$

Общий вывод, который можно сделать из этого примера, таков: взаимная теснота связи всех трех величин ($x=f_s$, $y=Q$ и $z=D$) достаточно велика, и можно, учитывая физический смысл явлений сделать заключение о ясной функциональной связи, выраженной соответствующими уравнениями.

4.3.5 Корреляционные (регрессионные) уравнения

Нахождение численной величины тесноты связи между аргументом и функциями (и также межфункциональной) является одной из основных задач корреляционного анализа. Другой основной задачей будет отыскание форм корреляционной связи в виде уравнений, что особо важно при многофакторном исследовании.

В последнем случае результат (функция) зависит от нескольких факторов (аргументов). Общий вид корреляционного (регрессионного) уравнения таков:

$$F_{xyz\dots} = f(x, y, z, \dots), \quad (4.16-4)$$

где $F_{xyz\dots}$ — значение результата, которое мы назовем общим функционалом;

x, y, z, \dots — значения факторов, предполагаемых в качестве независимых величин.

Связи в уравнении (4.16-4) должны быть однозначны, что требует подготовки таблиц опытных данных.

Можно показать, что уравнением типа (4.16-4) разрешается пользоваться лишь в том случае, когда сила связи между результатом и факторами не меньше, чем сила связи факторов между собой. Далее, на надежность результата F_{xyz} влияет надежность полученных значений факторов x, y, z, \dots

Принимая, что уравнение (4.16-4) удовлетворяет этим условиям, можно для алгебраического построения его пользоваться двухфакторными матрицами.

Вначале записывают (при постоянном шаге квантования главного фактора) дискретные значения каждого из факторов в каждом шаге и соответствующий результат. Общее количество записей должно не

только охватывать все особенности процесса, но и обеспечивать при классификации должную надежность при допустимой ошибке.

В соответствии с размахом колебаний значений каждого из факторов, записанные значения последних классифицируются и устанавливаются среднеклассовые значения, против которых записывают средний результат. В таблице 4.2-3 показан пример такой обработки синхронных записей по трем факторам и соответствующим значениям результата.

Таблица 4.2-3

Усредненные значения данных по факторам x, y, z и результата F_{xyz}

x	y	z	F_{xyz}	x	y	z	F_{xyz}
1	0,5	1	0,25	2	1,5	1	27,20
1	0,5	5	1,25	2	2,0	1	64,00
1	0,5	6	1,50	2	1,0	5	40,00
1	0,5	7	1,74	2	1,0	6	48,00
1	0,5	8	2,00	2	1,0	7	56,00
1	1,0	1	2,00	2	1,0	8	64,00
1	1,5	1	6,80	3	0,5	1	2,25
1	2,0	1	16,00	3	1,0	1	18,00
1	1,0	5	10,00	3	1,5	1	61,25
1	1,0	6	12,00	3	2,0	1	144,00
1	1,0	7	14,00	3	1,0	5	90,00
1	1,0	8	16,00	3	1,0	6	108,00
1	1,5	5	33,70	3	1,0	7	126,00
1	1,5	6	40,50	3	1,0	8	144,00
1	1,5	7	47,30	4	0,5	1	4,00
1	1,5	8	54,00	4	1,0	1	32,00
1	2,0	5	80,00	4	1,5	1	108,75
1	2,0	6	96,00	4	2,0	1	256,00
1	2,0	7	112,00	4	1,0	5	160,00
1	2,0	8	128,00	4	1,0	6	192,00
2	0,5	1	1,00	4	1,0	7	224,00
2	1,0	1	8,00	4	1,0	8	256,00

В этой по возможности ранжированной таблице каждая строка представляет средние значения по классам и результату. Составление такой таблицы — одна из наиболее трудоемких частей работы. Уменьшая эту подготовительную работу, например, за счет увеличения числа классов, мы неизбежно увеличиваем последующие расчеты, не говоря уже о возможном уменьшении надежности средних по классам.

Ранжирование по факторам позволяет находить корреляционные уравнения уже по таблицам, аналогичным таблице 4.2-3, если их данные позволяют отобрать группы значений факторов, в которых было бы нейтрализовано влияние всех факторов, за исключением одного. В этих целях отбирают группы чисел, в которых есть один переменный фактор при постоянстве остальных. Например, в таблице 4.2-3 этому соответствуют величины z при $x=y=1,0$ и результат изменяется только в функции величины z . Выписывая табличку

z	5	6	7	8
$F_{xy}z$	10	12	14	16

устанавливаем, что результат (назовем его функционалом F_z) можно выразить через величину z , или, иначе, $F_z = 2z$. Выписывая из таблицы 4.2-3 группу чисел при $x=2$ и $y=1$ и подставляя вместо значений z значения F_z , находим значение функционала F_{xz} . Наконец, при неизменных x и z (например, при $x=4$ и $z=1$), выполняя необходимые операции, находим общий функционал $F_{xyz} = 2x^2y^3z$. Отметим, что в некоторых случаях обобщенный коэффициент или (в случае полинома) свободный член уравнения можно найти только подбором по данным таблицы, аналогичной таблице 4.2-3 .

Но далеко не всегда по исходной таблице усредненных данных можно определить общий функционал $F_{xyz}...$ Поэтому, чтобы определить общий функционал $F_{xyz}...$, можно использовать матричный анализ. Для этого составляют двухфакторные матрицы. Далее в таблице 4.2-4 показана матрица x, x_j, y_i в которой каждому шагу каждого из факторов соответствует результат a_{ij} . В принципе матрицы прямоугольны, хотя из-за компактности изложения в примерах приводятся квадратные матрицы.

В предпоследних столбце и строке помещены суммы, по строкам $\sum_{i=1}^n a_{ij}$ и по столбцам $\sum_{j=1}^m a_{ij}$, как и сумма всех результатов $\sum_{i=1}^m \sum_{j=1}^n a_{ij}$. В последнем столбце помещены средние по строкам \bar{a}_i , а в последней строке — средние по столбцам \bar{a}_j , так же как и средняя из средних \bar{a}_{ij} , которую проще вычислить из соотношения:

$$\bar{a}_{ij} = \frac{\sum_{i=1}^m \sum_{j=1}^n a_{ij}}{m \cdot n} \tag{4.16-5}$$

Таблица 4.2-4

Двухфакторная матрица x_j, y_i

x_j	$j = 1$	\dots	$j = n$	$\sum_{i=1}^n a_{ij}$	\bar{a}_i
y_i					
$i = 1$	a_{11}	\dots	a_{1n}		
\dots					
$i = m$	a_{m1}	\dots	a_{mn}		
$\sum_{i=1}^m a_{ij}$				$\sum_{i=1}^m \sum_{j=1}^n a_{ij}$	
\bar{a}_j					\bar{a}_{ij}

Двухфакторная матрица указанной в таблице 4.2-4 формы имеет ту особенность, что в ней ряды значений факторов усреднены, и поэтому ряды величин \bar{a}_i и \bar{a}_j в наибольшей возможной степени взаимно нейтральны и отображают дискретные значения зависимости результата от дискретных значений каждого из факторов. Все это позволяет найти уравнения рядов \bar{a}_i и \bar{a}_j ; эти ряды при необходимости сглаживают одним из способов. Назовем зависимость ряда \bar{a}_j от ряда x_j , выраженную уравнением, первичным функционалом F_x , а соответствующую зависимость ряда \bar{a}_i от ряда y_i — первичным функционалом F_y . Разумеется, при числе матриц больше одной коды функционалов могут изменяться в соответствии с кодом факторов.

Число матриц для получения корреляционного уравнения равно числу сочетаний из K факторов по 2. Поэтому для составления уравнения по двум факторам нужна одна матрица, по трем факторам — три матрицы, по четырем — шесть матриц и т. д.

Приведем краткие правила вывода корреляционных уравнений и два примера для простых задач, решаемых без помощи ЭВМ.

1. Сначала определяют по всем матрицам первичные функционалы. Они могут быть или одночленами, или полиномами.

2. Затем определяют матричные функционалы (F_{xy} или F_{yz} и т.д.):

а) если первичные функционалы данной матрицы представлены одночленами, то для получения матричного функционала их произведение делится на среднюю из средних. Иначе, матричный функционал факторов x и y

$$F_{xy} = \frac{F_x F_y}{\bar{a}_{ij}}; \quad (4.16-6)$$

б) если первичные функционалы представлены полиномами, то для получения матричного функционала их алгебраически суммируют и из этой суммы вычитают среднюю из средних. Иначе, в этом случае

$$F_{xy} = F_x + F_y - \bar{a}_{ij}; \quad (4.16-7)$$

в) если один первичный функционал представлен одночленом, а другой — полиномом, то один из них следует преобразовать так, чтобы в результате можно было пользоваться формулами (4.16-6) или (4.16-7).

Отметим, что получение первичных и матричных функционалов позволяет установить принципиальную структуру общей корреляционной формулы. Вспомогательные рисунки кривых функционалов также могут служить этой цели.

3. По матричным функционалам устанавливают общий функционал всех факторов $F_{xyz...}$ в виде корреляционного уравнения.

Для уяснения последующего принимают во внимание, что в матричных функционалах символы факторов могут повторяться в различных сочетаниях, в различных степенях и с различными коэффициентами. Поэтому действия по составлению корреляционного уравнения можно скорее назвать сопоставлением матричных функционалов с использованием своеобразных матричных норм, выбираемых из уравнений матричных функционалов:

а) если все матричные функционалы представлены одночленами, общий функционал (назовем его общим функционалом первого рода) может быть выражен произведением факторных символов высших степеней, взятых с соответствующими коэффициентами; если в уравнениях матричных функционалов есть одинаковые символы одной и той же степени, для произведения берут один из них. Таким образом, в этом случае

$$F'_{xyz...} = ax^{\max} y^{\max} z^{\max} \dots, \quad (4.16-8)$$

где a —общий коэффициент, а верхними индексами обозначена указанная выборка из уравнений матричных функционалов;

б) если все матричные функционалы представлены полиномами, то общий функционал (назовем его функционалом второго рода) может быть представлен в виде алгебраической суммы символов факторов высших степеней с их коэффициентами и свободного члена матричных уравнений; если есть одинаковые символы одной и той же степени, для суммирования берется один из них. То же относится к свободному члену.

В данном случае

$$F_{xyz\dots}^n = ax^{\max} + by^{\max} + cz^{\max} + \dots + d. \quad (4.16-9)$$

Заполним на основании данных таблицы 4.2-3 матрицу $xу$, что представлено таблицей 4.2-5, причем тут возможны два варианта:

а) если по исходной таблице усредненных данных имеется возможность заполнить матрицу с нейтрализацией всех остальных факторов (кроме необходимых двух, в данном случае x, y), то в большинстве случаев это дает более точное решение;

б) если такой возможности нет, в матричные клетки a_{ij} записывают все данные по соответствующим результатам, в суммирующих клетках — все суммы, в столбце \bar{a}_i и строке \bar{a}_j — средние по всем записям, получая, таким образом, среднему взвешенные значения результатов от данного фактора.

Таблица 4.2-5

Матрица $xу$ (при $z=1$)

y	x					Σ	\bar{a}_i
	1	2	3	4			
0,5	0,25	1,00	2,25	4,00	7,50	1,88	
1,0	2,00	8,00	18,00	32,00	60,00	15,00	
1,5	6,80	27,20	61,25	108,75	204,00	51,00	
2,0	16,00	64,00	144,00	256,00	480,00	120,00	
Σ	25,05	100,20	225,50	400,75	751,50		
\bar{a}_j	6,26	25,05	56,40	100,19		47,00	

В таблице 4.2-5 возможности заполнения матрицы даны значениями результата при $z=1$.

Отсюда первичные функционалы (получены почти без сглаживания)

$$F_x = 6,26x^2 \text{ и } F_y = 15y^3.$$

На основании формулы (4.16-6) получаем:

$$F_{xy} = \frac{6,26 x^2 \cdot 15y^3}{47} = 2x^2 y^3.$$

Если записать в матрицу $xу$ все результаты таблицы 4.2-3 получим ряды (самой матрицы для краткости не приводим):

\bar{y}	0,5	1,0	1,5	2,0
\bar{a}_i	1,75	80,65	47,40	112,00
x	1	2	3	4
\bar{a}_j	33,7	37,6	86,7	154,3

При общей средней, вычисленной по формуле (4.16-5), $\bar{a}_{ij}=66,00$. Как видно, кривые значений \bar{a}_i и \bar{a}_j требуют тщательного сглаживания, особенно выделяются $\bar{a}_i=80,65$ и $\bar{a}_j=33,7$. Для улучшения сглаживания введем среднюю \bar{a}_{ij} , соответствующую $y=1,25$ и $x=2,50$. После сглаживания получаем:

$$F_x = 9,5x^2 \text{ и } F_y=14y^3.$$

Тогда

$$F_{xy} = \frac{6,26 x^2 \cdot 15y^3}{47} = 2x^2 y^3.$$

На практике могут быть клетки матрицы с прочерками (пустые). При тщательно поставленных опытах и заполнении начальных, средних и конечных строк и столбцов могут быть получены удовлетворительные результаты, что может сказаться и на определении количества опытов при их планировании. Это лучше делать (при возможности заполнения вслед за каждым опытом с синхронной записью) по нескольким факторам подготовленной классификации (если знают размах колебаний факторов) и проведения последовательного анализа.

В таблицах 4.2-6 и 4.2-7 показаны матрицы xz и yz , причем результаты при $z= 1$ опущены, чтобы не нарушать постоянства шага опытных данных.

Таблица 4.2-6

Матрица xz

z	x				Σ	\bar{a}_i
	1	2	3	4		
5	1,25; 10,00; 33,70; 80,00	40,00	90,00	160,00	414,95	59,3
6	1,50; 12,00; 40,50; 96,00	48,00	108,00	192,00	498,00	71,2
7	1,74; 14,00; 47,30; 112,00	56,00	126,00	224,00	581,04	83,00
8	2,00; 16,00; 54,00; 128,00	64,00	144,00	256,00	664,00	94,8
Σ	649,99	208,00	468,00	832,00	2157,99	
\bar{a}_i	40,70	52,00	117,00	208,00		76,30

Сглаживая, получаем равенства $F_x = 13 x^2$ (с наибольшей ошибкой для \bar{a}_{j1}) и $F_z = 11,9 z$. Тогда

$$F_{xz} = \frac{13 \cdot 11,9 x^2 z}{76,3} = 2x^2 z.$$

По сглаживанию рядов \bar{a}_i и \bar{a}_j получаем:

$$F_y = 13 y^3; F_z = 6,25 z; F_{yz} = \frac{13 \cdot 6,25 y^3 z}{40,7} = 2y^3 z.$$

Сопоставляя матричные функционалы, получаем:

$$F_{xyz} = 2x^2 y^3 z.$$

Таблица 4.2-7

Матрица yz

z	y				Σ	\bar{a}_i
	0,5	1,0	1,5	2,0		
5	1,25	10,00	33,70	80,00	124,95	31,24
6	1,50	12,00	40,50	96,00	150,00	37,50
7	1,74	14,00	47,30	112,00	175,04	43,76
8	2,00	16,00	54,00	128,00	200,00	50,00
Σ	6,49	52,00	175,00	416,00	649,99	
\bar{a}_i	1,62	13,00	43,90	104,00		40,70

Эта структура уравнения повторяет структуру формулы (4.16-8).

Объективное установление структуры корреляционного уравнения и степенных зависимостей — шаг на пути правильного установления и рациональной интерпретации закона развития явления. Эта структура может быть, конечно, установлена и логическим путем, если в ее основу положены известные законы развития явления.

Второй пример несколько сокращен: опущена исходная таблица усредненных результатов, аналогичная таблице 4.2-3 и часть пояснений. Этот пример представлен таблицами 4.2-8—4.2-10.

Таблица 4.2-8.

Матрица xy

y	x					Σ	\bar{a}_t
	1	2	3	4			
0,5*	-3,37	-0,38	4,63	11,62	12,50	3,13	
1,0	-2,50	0,50	5,50	12,50	16,00	4,00	
1,5	0,13	2,88	7,87	14,87	25,75	6,44	
2,0	4,50	7,50	12,50	19,50	44,00	11,00	
Σ	-1,24	10,50	30,50	58,49	98,25		
\bar{a}_j	-0,31	2,62	7,62	14,62		6,15	

Рассматривая ряды средних, можно установить, что они не являются просто степенными функциями, а включают еще и второй одночлен. В самом деле, после сглаживания $F_x = x^2 - 1,38$ и $F_y = y^3 + 3$, так что соответственно формуле (4.16-7)

$$F_{xy} = x^2 + y^2 - 1,38 + 3 - 6,15 = x^2 + y^3 - 4,53.$$

Таблица 4.2-9.

Матрица xz

z	x					Σ	\bar{a}_t
	1	2	3	4			
5	1,12	4,13	9,12	16,13	30,50	7,63	
6	0,13	3,12	8,13	15,12	26,50	6,63	
7	-0,89	2,13	7,12	14,13	22,49	5,62	
8	-1,88	1,12	6,13	13,12	18,49	4,62	
Σ	-1,52	10,50	30,50	58,50	97,98		
\bar{a}_j	-0,38	2,63	7,63	14,63		6,12	

Таблица 4.2-10.

Матрица zy

z	y					Σ	\bar{a}_i
	0,5	1,0	1,5	2,0			
5	4,63	5,50	7,87	12,50	30,50	7,63	
6	3,63	4,50	6,87	11,50	26,50	6,63	
7	2,63	3,50	5,87	10,50	22,50	5,63	
8	1,63	2,50	4,87	9,50	18,50	4,63	
Σ	12,52	16,00	25,48	44,00	98,00		
\bar{a}_j	3,13	4,00	6,37	11,00		6,13	

В этом случае общая структура корреляционного уравнения близка к формуле (4.26-9).

Отсюда

$$F_x = x^2 - 1,37; F_z = 12,63 - z.$$

По формуле (4.26-7)

$$F_{xz} = x^2 - z + 12,63 - 1,37 - 6,12 = x^2 - z + 5,14.$$

Отсюда

$$F_y = y^3 + 3; F_z = 12,63 - z; F_{yz} = 12,63 + 3 - 6,13 + y^3 - z = y^3 - z + 9,50.$$

Сопоставляя матричные функционалы, находим, что общий вид корреляционного уравнения будет:

$$F_{xyz} = x^2 + y^3 - z + d.$$

Значение d определяем подбором; в данном случае $d=2$.

Описанные примеры приведены для расчетов вручную; но все изложенное имеет самостоятельное методическое значение.

В приложениях даны формы и аналитические выражения для встречающихся простых функционалов. В более сложных случаях исходную таблицу дискретных синхронных измерений, ранжированную по факторам, для получения корреляционного уравнения обрабатывают на ЭВМ. При использовании ЭВМ все расчеты обычно разбивают на три части: вычисление исходной таблицы (примерно по форме таблицы 4.2-3); определение принципиальной структуры корреляционного уравнения и установление численных значений показателей степеней, обобщенных коэффициентов и свободных членов.

Отметим, что если в какой-либо матрице ряд \bar{a}_i или \bar{a}_j имеет одинаковые (или почти одинаковые) значения, то соответствующий фактор не влияет на значение матричного функционала, и он из

последнего может быть исключен.

В некоторых случаях простой корреляционной зависимости, как это показано в примерах, мы не получим, так как явление осложняется влиянием неучитываемых факторов (например, времени, химического состава обрабатываемой среды и др.). В таких случаях рекомендуется сначала определить структуру многофакторного регрессионного уравнения и лишь после этого уточнить входящие в состав уравнения члены, чтобы получить уравнение с наименьшей ошибкой. Структуру уравнения определяют либо логически (если общие законы взаимосвязи факторов известны), либо при помощи приближенного представления рядов средних \bar{a}_i и \bar{a}_j .

Первый способ не требует пояснений. При втором способе лучше построить графики первичных функционалов и по ним определить приближенные (с ошибкой) простые зависимости. Предположим, что при трех факторах x , y , z , эти зависимости дают функционалы

$$F_x = ax^k, F_y = by^l \text{ и } F_z = \frac{c}{z^m},$$

— то есть грубо взятый функционал

$$F_x = ax^k, F_y = by^l \text{ и } F_z = \frac{c}{z^m},$$

Ценность предварительного определения структуры уравнения заключается в приближении его к рациональной форме. Затем уточняют значения каждого из функционалов так, чтобы получить минимальную ошибку.

Пусть, например, с допустимой ошибкой

$$F_x = a_1(x^k - a_2x + g_1); F_y = b_1(y^{0,8l} + b_2y - g_2) \text{ и } F_z = \frac{c}{z^{0,9m}}.$$

Тогда с приемлемой максимальной ошибкой (которую для опытных пределов следует указать)

$$F_{xyz} = d \frac{(x^k - a_2x + g_1)(y^{0,8l} + b_2y - g_2)}{z^{0,9m}}.$$

4.3.6. Ранговая корреляция

Задачей метода ранговой корреляции является оценка связи между упорядоченными объектами исследования. Для упорядочения объектов используются ранги и тогда задача состоит в том, чтобы по совокупности данных экспертами упорядочений составить усредненное упорядочение, наиболее близкое к истинному.

К решению задачи применяется статистический подход. Каждая данная экспертом ранжировка считается несовершенным, искаженным субъективными воздействиями вариантом правильного

упорядочения объектов. Тогда к оценке полученных ранжировок можно применять статистические критерии, в частности, для согласованности мнений экспертов используется коэффициент конкордации (согласия), который меняется от нуля (нет согласия) до единицы (полное согласие).

Метод ранговой корреляции.

Пусть имеются объекты x_1, x_2, \dots, x_n . Специалистам, хорошо знакомым с исследуемыми объектами, предлагается расположить объекты, предположим, в порядке убывания степени их влияния на некоторый показатель, т. е. присвоить ранг r_{ij} — ранг для каждого j -го объекта, данный i -ым экспертом.

Пояснение. Сводная матрица $[r_{ij}]$ размерами $m \times n$ содержит по строкам натуральный ряд чисел, расположенных в различном порядке. Среднее арифметическое натурального ряда чисел равно

$$\frac{1}{n} (1 + 2 + \dots + n) = \frac{n + 1}{2}$$

и тогда общее среднее равно

$$a = \frac{1}{2} m(n + 1).$$

В матрице $[r_{ij}]$ возможны и дробные ранги. Они появляются тогда, когда эксперт не может отдать предпочтение двум или большему числу объектов, например, если эксперт не может упорядочить объекты с рангами 4 и 5, то он может приписать этим двум объектам ранг $(4+5)/2=4,5$.

План эксперимента представляет собой анкеты, получаемые экспертами. Суммарная анкета — таблица, по строкам которой располагаются ранжировки экспертов, а по столбцам — ранги объектов исследования (табл. 4.1-А).

Расчет средних:

$$a_n = \frac{n + 1}{2}$$

— среднее ранжировочного ряда эксперта,
где n — число объектов исследования;

$$a = \frac{1}{2} m(n + 1)$$

— общее среднее таблицы рангов,
где m — число экспертов.

Таблица 4.1-А.

Результаты расчетов методом ранговой корреляции

$i \backslash j$	Объекты исследования			
	x_1	x_2	...	x_n
Номер эксперта				
1	r_{11}	r_{12}	...	r_{1n}
2	r_{21}	r_{22}	...	r_{2n}
...
m	r_{m1}	r_{m2}	...	r_{mn}
Сумма рангов	$\sum_{i=1}^m r_{i1}$	$\sum_{i=1}^m r_{i2}$...	$\sum_{i=1}^m r_{in}$

Расчет суммы квадратов отклонений. Первоначально рассчитываются квадраты отклонений суммарных рангов от общего среднего:

$$d_j^2 = \left(\sum_{i=1}^m a_{ij} - a \right)^2,$$

затем суммы квадратов отклонений:

$$S(d_j^2) = \sum_{j=1}^n \left(\sum_{i=1}^m a_{ij} - a \right)^2.$$

Расчет максимальной суммы квадратов отклонений. Доказано, что при полном согласии мнений экспертов максимальная сумма квадратов будет равна:

$$S(d_j^2) = \frac{1}{12} m^2 (n^3 - n).$$

Расчет показателя дробных рангов. При наличии дробных рангов максимальная сумма квадратов уменьшается на

$$m \sum_{i=1}^e T_i,$$

где

$$T_i = \frac{1}{12} \sum_{v=1}^g (t_v^3 - t_v);$$

e — число строк матрицы, содержащей связанные ранги; g — число типов связанных рангов в строке; t — количество равных рангов в i -ой строке.

Расчет коэффициента конкордации. При отсутствии связанных рангов коэффициент конкордации получается как отношение рассчитанной $S(d_j^2)$ к максимальной $S(d_j^2)_{\max}$:

$$V_p = \frac{S(d_j^2)}{S(d_j^2)_{\max}} = \frac{\sum_{j=1}^n \left(\sum_{i=1}^m a_{ij} - a \right)^2}{\frac{1}{12} m^2 (n^3 - n)};$$

при наличии связанных рангов

$$W_p = \frac{\sum_{j=1}^n \left(\sum_{i=1}^m a_{ij} - a \right)^2}{\frac{1}{12} m^2 (n^3 - n) - m \sum_{i=1}^e T_i}.$$

Оценка коэффициента конкордации. Установлено, что для числа объектов $n > 7$ величина $[m(n-1)W_p]$ имеет χ^2 -распределение (при $n < 7$ используются специальные таблицы Кендалла):

$$\chi_p^2 = m(n-1)W_p.$$

Принятие решений. Если расчетное значение χ^2 -распределения окажется меньше табличного

$$\chi_p^2 < \chi_r^2 (f = n - 1, q = 0,05)$$

при числе степеней свободы f и заданном уровне значимости q , то мнения экспертов считают согласованными. В этом случае строят диаграмму рангов (гистограмму), где по оси абсцисс дискретно откладывают объекты ранжирования, а по оси ординат суммы рангов в обратном порядке. Вид сглаженной ранжировочной кривой в какой-то мере позволяет решить задачу классификации объектов (при двух классах).

Если приведенное выше неравенство не выполняется, т. е. мнения экспертов не согласованы, то одним из решений может быть изменение состава экспертов или их обучение. Иногда меняют и количество объектов.

Примечание. Кроме ранжирования объектов в некоторых случаях приходится с целью разработки весов ранжировать и качественные показатели объектов. При этом метод усложняется.

4.4. Дисперсионный анализ

Задачей дисперсионного анализа является исследование влияния различных факторов на изменчивость средних значений наблюдаемых случайных величин. План дисперсионного анализа предполагает проведение эксперимента, позволяющее разложить сложную дисперсию на составляющие (при выполнении гипотезы об аддитивности дисперсий изучаемой случайной величины). Далее полученные дисперсии оцениваются по определенному критерию. Наиболее удобно и оправдано сравнивать полученные дисперсии с ошибкой опытных данных, которую получают по параллельным опытам на каждом (или на некоторых) уровне факторов.

4.4.1. Однофакторный дисперсионный анализ.

Необходимо выяснить влияние фактора A на выходную переменную y . Предполагается, что результаты наблюдений можно представить моделью

$$\hat{y}_{ij} = \bar{y} + \alpha_i + e_{ij} \quad (j = 1, 2, \dots, m, i = 1, 2, \dots, p), \quad (4.17)$$

где \bar{y} — суммарный эффект во всех опытах (среднее значение); α_i — эффект фактора A на i -ом уровне; e_{ij} — ошибка измерения на i -ом уровне; m — число параллельных опытов; p — число уровней фактора A . Очевидно, общее число опытов

$$N = \sum_{i=1}^p m_i \quad (4.18)$$

или при одинаковом числе параллельных опытов

$$N = pm. \quad (4.19)$$

Предполагается, что закон распределения случайной величины нормальный, а рассеяние опытов на каждом уровне фактора A не превышает некоторой величины или дисперсии на каждом уровне однородны.

План эксперимента и подготовительные расчеты представлены в табл. 4.2.

Расчет средних:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^{m_i} y_{ij} \quad \text{общее среднее;} \quad (4.20)$$

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \quad \text{среднее по уровням фактора } A \text{ (строкам).} \quad (4.21)$$

Расчет сумм квадратов отклонений. Сумма квадратов отклонений, связанная с рассеиванием относительно общего среднего:

$$S_{\text{ост}} = \sum_{i=1}^p \sum_{j=1}^{m_i} (y_{ij} - \bar{y})^2; \quad (4.22)$$

сумма квадратов отклонений, связанная с рассеиванием между уровнями фактора A (междууровневая):

$$S_A = \sum_{i=1}^p m_i (y_i - \bar{y})^2; \quad (4.23)$$

сумма квадратов отклонений, связанная с рассеиванием внутри выборки по уровням (внутриуровневая):

$$S_R = S_{\text{ост}} - S_A; \quad (4.24)$$

сумма квадратов отклонений, связанная с рассеиванием по параллельным опытам на i -ом уровне:

$$S_i = \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2. \quad (4.25)$$

Таблица 4.2.

Результаты расчетов при однофакторном дисперсионном анализе

План		Результаты расчетов		
i	j	Параллельные опыты 1, 2, . . . , m_i	Суммы результатов по параллельным опытам	Средний по параллельным опытам
Уровни фактора A	(1)	$y_{11}y_{12} \dots y_{1m_1}$	m_1 $\sum_{j=1}^{m_1} y_{1j}$	\bar{y}_1
	(2)	$y_{21}y_{22} \dots y_{2m_2}$	m_2 $\sum_{j=2}^{m_2} y_{2j}$	\bar{y}_2
	⋮	⋮	⋮	⋮
	⋮	⋮	⋮	⋮
	p	$y_{p1}y_{p2} \dots y_{pm_p}$	m_p $\sum_{j=1}^{m_p} y_{pj}$	\bar{y}_p

Расчет дисперсий. Рассчитанные суммы имеют соответственно следующие степени свободы f :

$$f_R = N - p = p(m - 1); \quad (4.26)$$

$$f_{\text{ост}} = N - 1; \tag{4.27}$$

$$f_A = p - 1 \tag{4.28}$$

$$f_i = m_i - 1. \tag{4.29}$$

Из условия (4.24)

$$N - p = N - 1 - (p - 1),$$

т. е. получается тождество.

Рассчитаем дисперсии:

$$s_A^2 = S_A / f_A \text{ — дисперсия фактора } A; \tag{4.30}$$

$$s_{\text{ост}}^2 = S_{\text{ост}} / N - 1 \text{ — остаточная дисперсия}; \tag{4.31}$$

$$s_R^2 = S_R / N - p \text{ — ошибка опыта (дисперсия воспроизводимости)}; \tag{4.32}$$

$$s_i^2 = S_i / f_i \text{ — дисперсия параллельных опытов на } i\text{-ом уровне}. \tag{4.33}$$

Проверка однородности дисперсий. Однородность дисперсий на каждом уровне фактора A можно проверить по критерию Кохрена, расчетное значение которого определяют по уравнению:

$$G = s_{i \max}^2 / \sum_{i=1}^p s_i^2, \tag{4.34}$$

где $s_{i \max}^2$ — максимальная из рассчитанных дисперсий параллельных опытов (построчных дисперсий, см. табл. 4.2); $\sum_{i=1}^p s_i^2$ — сумма всех дисперсий по уровням фактора A .

Если выполняется условие

$$G_p < G_T \quad (f_i = m_i - 1, \quad f_2 = p, \quad q = 0,05), \tag{4.35}$$

то гипотеза об однородности дисперсий правомерна G_T находят по таблице критерия Кохрена для степеней свободы f_1 (максимальная дисперсия), f_2 (число уровней) и заданного уровня значимости q . В технических расчетах принят 5%-ный уровень значимости.

Оценка различия дисперсий. Строится отношение дисперсии фактора A к ошибке опыта и получается расчетное значение критерия Фишера F_p :

$$\frac{s_A^2}{s_R^2} = F_p. \tag{4.36}$$

Принятие решений. Если расчетное значение критерия Фишера

$$F_p > F_T \quad (f_A = p - 1, \quad f_R = p(m - 1), \quad q = 0,05) \tag{4.37}$$

(F_T — табличное значение критерия Фишера для степеней свободы f_A , f_R и заданного уровня значимости q) то влияние уровней фактора A существенно. Условие (4.36) означает, что влияние фактора A , разделенного по уровням превышает уровень ошибки опытных данных.

Результаты однофакторного дисперсионного анализа удобно представить в виде табл.4.3, где расчет производился по преобразованным формулам (4.22), (4.23) и (4.24).

Таблица 4.3.

Результаты расчетов при однофакторном дисперсионном анализе

Источник изменчивости	Расчетные формулы и значения			
	S	f	s^2	F_p
Между уровнями фактора	$S_A = \sum_{i=1}^p m_i \bar{y}_i^2 - N \bar{y}^2$	$p - 1$	$s_A^2 = \frac{S_A}{p - 1}$	$\frac{s_A^2}{s_R^2}$
Внутри уровней фактора	$S_R = \sum_{i=1}^p \sum_{j=1}^{m_i} y_{ij}^2 - \sum_{i=1}^p m_i \bar{y}_i^2$	$p(m-1)$	$s_R^2 = \frac{S_R}{p(m-1)}$	
	$S_{ост} = \sum_{i=1}^p \sum_{j=1}^{m_i} y_{ij}^2 - N \bar{y}^2$	$N - 1$		

Пример1. Рассмотрим задачу исследования активности катализатора в зависимости от способа его получения. Катализатор для химической реакции получали четырьмя различными способами, которые соответствовали $p = 4$ уровням фактора A . В эксперименте измерялась активность катализатора, которая выражалась некоторой величиной y . Для каждого уровня было сделано по одинаковому числу ($m_i=5$) измерений. Необходимо проверить независимость качества катализатора (его активности) от способа получения. Измерение значения активности катализатора в некоторых условных единицах приведены в табл. 4. 4.

Таблица 4.4.

Результаты эксперимента при однофакторном дисперсионном анализе

i	Параллельные опыты					Σ	\bar{y}_i
	1	2	3	4	5		
Уровни фактора А							
1	56	55	62	59	60	292	58,4
2	64	61	50	55	56	286	57,2
3	45	46	45	39	43	218	43,6
4	42	39	45	43	41	210	42,0
						$\Sigma\Sigma = 1006$	$\bar{y} = 50,3$

В этой же таблице приведены средние, рассчитанные по формулам (4.20) и (4.21). В табл. 4.5 приведены результаты расчетов для оценки однородности дисперсий, рассчитанные по формулам (4.25) и (4.33).

Таблица 4.5.

Расчеты для оценки однородности дисперсий

i	Параллельные опыты					S_i	s_i^2 ($f_i = 3$)
	1	2	3	4	5		
Уровни фактора А							
1	2,4 ²	3,4 ²	3,6 ²	0,6 ²	1,6 ²	33,20	11,07
2	6,8 ²	3,8 ²	7,2 ²	2,2 ²	1,2 ²	118,80	39,60
3	1,4 ²	2,4 ²	1,4 ²	4,6 ²	0,6 ²	31,20	10,40
4	0 ²	3 ²	3 ²	1 ²	1 ²	20,00	6,66
						$\bar{S}_R = 203,20$	$\Sigma = 67,72$

Если учесть, что $s_{i \max}^2 = 39,60$, то по формуле (4.34)

$$G_p = \frac{39,60}{67,72} = 0,585,$$

и по условию (4.35) (см. приложение 7)

$$G_p < G_\tau = 0,6841 \quad (f_1 = 4 - 1 = 3, \quad f_2 = 4, \quad q = 0,05).$$

Вывод: дисперсии однородны.

Результаты расчета сумм квадратов отклонений и дисперсий по формулам (4.22) - (4.32) сведены в табл. 4. 6.

Таблица 4.6.

Результаты расчетов при однофакторном дисперсионном анализе

Источник изменчивости	Расчетные формулы и значения			
	<i>S</i>	<i>f</i>	<i>s</i> ²	<i>F</i>
Между уровнями фактора Внутри уровней фактора	$S_A = 1135,0$	$f_A = 4 - 1 = 3$	$s_A^2 = 378,3$	$F_D = 29,8$
	$S_R = 203,2$	$f_R = 4(5 - 1) = 16$	$s_R^2 = 12,7$	
	$S_{\text{ост}} = 1338,2$	$f_{\text{ост}} = 4 \cdot 5 - 1 = 19$		

Расчетное значение критерия Фишера оказывается большим табличного (см. приложение б):

$$F_D = 29,8 > F_T = 3,24 \quad (f_A = 3, f_R = 16, q = 0,05).$$

Таким образом, влияние фактора *A* превышает уровень ошибок опытных данных, или на качество катализатора влияют способы его получения.

4.4.2. Двухфакторный дисперсионный анализ.

Необходимо выяснить влияние двух факторов на некоторую выходную переменную (функцию отклика). Предполагается, что исследуется влияние факторов *A* и *B*, варьируемых на *p*- и *k*-ом уровнях соответственно. Математическая модель эксперимента следующая:

$$y_{ij} = \bar{y} + \alpha_i + \beta_j + e_{ij} \quad (i = 1, 2, \dots, p, j = 1, 2, \dots, k), \quad (4.38)$$

где α_i и β_j — эффекты факторов *A* и *B* соответственно на *i*- и *j*-ом уровнях; \bar{y} — суммарный эффект факторов во всех опытах; e_{ij} — ошибка измерения (помеха), имеющая нормальное распределение, нулевое математическое ожидание и определенную дисперсию. Общее число опытов при однократных наблюдениях

$$N = p - k. \quad (4.39)$$

План эксперимента и некоторые расчеты сумм и средних представлены в табл. 4.7.

Расчет средних. Можно получить следующие выборочные средние:

$$\bar{y}_i = \frac{1}{k} \sum_{j=1}^k y_{ij} \quad \text{— среднее по строкам;} \quad (4.40)$$

$$\bar{y}_j = \frac{1}{p} \sum_{i=1}^p y_{ij} \quad \text{— среднее по столбцам;} \quad (4.41)$$

$$\bar{y} = \frac{1}{pk} \sum_{i=1}^p \sum_{j=1}^k y_{ij} \quad \text{— общее среднее.} \quad (4.42)$$

Таблица 4.7.

План эксперимента и расчеты при двухфакторном дисперсионном анализе

$i \backslash j$	Уровни фактора B 1, 2, . . . , k	Средние по строкам
Уровни фактора A		
1	$y_{11}y_{12}, \dots, y_{1k}$	\bar{y}_1
2	$y_{21}y_{22}, \dots, y_{2k}$	\bar{y}_2
.
.
p	$y_{p1}y_{p2}, \dots, y_{pk}$	\bar{y}_p
Средние по столбцам	$\bar{y}_1\bar{y}_2, \dots, \bar{y}_k$	Общее среднее

Расчет сумм квадратов отклонений:

$$S_A = k \sum_{i=1}^p (\bar{y}_i - \bar{y})^2; \quad (4.43)$$

$$S_B = p \sum_{j=1}^k (\bar{y}_j - \bar{y})^2; \quad (4.44)$$

$$S_{\text{ост}} = \sum_{i=1}^p \sum_{j=1}^k (y_{ij} - \bar{y})^2; \quad (4.45)$$

$$S_R = S_{\text{ост}} - S_A - S_B = \sum_{i=1}^p \sum_{j=1}^k (y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2. \quad (4.46)$$

Расчет дисперсий. Рассчитанные суммы S_A , S_B , S_R имеют следующие степени свободы:

$$f_A = p - 1; \quad (4.47)$$

$$f_B = k - 1; \quad (4.48)$$

$$f_R = (p-1)(k-1). \quad (4.49)$$

Тогда дисперсии можно рассчитывать по формулам:

$$s_A^2 = S_A/f_A; \quad (4.50)$$

$$s_B = S_B/f_B; \quad (4.51)$$

$$s_R^2 = S_R/f_R. \quad (4.52)$$

Оценка отношений дисперсий. Строятся отношения дисперсий факторов A и B к ошибке опыта и получаются расчетные значения критерия Фишера:

$$F_{pA} = \frac{s_A^2}{s_R^2}; \quad F_{pB} = \frac{s_B^2}{s_R^2}. \quad (4.53)$$

Принятие решений. Если расчетные значения критерия Фишера (F_T — табличное значение критерия Фишера для степеней свободы f_A или f_B заданного уровня значимости q), то влияние факторов A и B на выходную переменную существенно (при выбранных p - и k -ом уровнях).

Результаты и расчеты двухфакторного дисперсионного анализа удобно представить в виде табл. 4.8.

Примечание. Линейная модель (4.38) справедлива, если между факторами A и B отсутствует взаимодействие, т. е. дисперсия $s_{AB}^2 = 0$. Если такое взаимодействие имеется, то s_{AB} входит в дисперсию ошибки опыта s_R^2 и выделить ее можно только при наличии параллельных опытов по уровням факторов A и B . Такой алгоритм хотя принципиально и не отличается от приведенного выше, но имеет особенности.

Таблица 4.8.

Результаты расчетов при двухфакторном дисперсионном анализе

Источник изменчивости	Расчетные формулы и значения			
	S	f	s^2	F
Фактор A	$S_A = k \sum_{i=1}^p \bar{y}_i^2 -$ $= pky^2$	$p - 1$	s_A^2	$\frac{S_A^2}{S_R^2} = F_{pA}$
Фактор B	$S_B = p \sum_{j=1}^k \bar{y}_j^2 -$ $= pky^2$	$k - 1$	s_B^2	$\frac{S_B^2}{S_R^2} = F_{pB}$
Ошибка опыта	$S_R = \sum_{i=1}^p \sum_{j=1}^k y_{ij}^2 -$ $= k \sum_{i=1}^p \bar{y}_i^2 -$ $= p \sum_{j=1}^k \bar{y}_j^2 + pky^2$	$(p-1)(k-1)$	$s_R^2 = s_0^2$	
	$S_{\text{ост}} = S_A + S_B +$ $+ S_R$	$pk - 1$		

Пример 2. Исследовать конверсию процесса радикальной полимеризации. Исследовалось влияние на процесс радикальной полимеризации двух факторов: A — тип растворителя на уровнях (a_1, a_2, a_3, a_4) и B — тип галоидных алкилов (b_1 — CH_3 , b_2 — C_3H_7 , b_3 — C_4H_9 , b_4 — $\text{C}_2\text{H}_5\text{Br}$). Выход полимера в процентах представлен в табл. 4.9.

Таблица 4.9.

Результаты эксперимента при двухфакторном дисперсионном анализе

	Уровни фактора B				Средние по строкам
	b_1	b_2	b_3	b_4	
Уровни фактора A					
a_1	13,2	18,9	7,3	20,0	$59,4 : 4 = 14,85$
a_2	4,7	19,8	38,2	60,1	$122,8 : 4 = 32,20$
a_3	53,4	14,0	5,1	19,6	$92,1 : 4 = 23,02$
a_4	13,6	9,5	54,4	58,2	$135,7 : 4 = 33,92$
	$\frac{84,9}{4} = 21,2$	$\frac{62,2}{4} = 15,6$	$\frac{105,0}{4} = 26,3$	$\frac{157,9}{4} = 39,5$	$\frac{410,0}{16} = 25,6$

Рассчитанные по формулам (4.40), (4.41), (4.41) средние и формулы расчета сумм квадратов отклонений (4.43), (4.44), (4.45), использованы для заполнения табл. 4.10.

Таблица 4.10.

Расчеты по схеме двухфакторного анализа

Источники изменчивости	Расчетные формулы и значения			
	S	f	s^2	F
Фактор A	$S_A = 1273,00$	$f_A = 3$	424,33	$F_{pA} = 1,34$
Фактор B	$S_B = 1994,00$	$f_B = 3$	664,66	$F_{pB} = 2,11$
Ошибка опыта	$S_R = 2829,00$	$f_R = 9$	314,33	$F_T = 3,86$
	$S_{ост} = 6096,22$	$f_{ост} = 15$		

Отсюда можно сделать вывод, что при уровне значимости $q = 0,05$ ($F_{pA} < F_T$ и $F_{pB} < F_T$) факторы A и B не влияют на выход полимера. Если исследователи сомневаются в правильности сделанного вывода, то можно дополнительно провести опыты и, изменив схему дисперсионного анализа, произвести новые расчеты. Для рассмотренной здесь задачи проведены параллельные опыты и окончательный результат изменился — факторы A и B оказались значимыми. По-видимому, невысокая воспроизводимость опытов сильно влияет на результат исследования.

4.4.3. Дисперсионный анализа при использовании схемы латинского квадрата.

В предыдущих методах дисперсионного анализа число необходимых опытов определялось произведением уровней исследуемых факторов. Если предположить, что число уровней одинаково, то в двухфакторном дисперсионном анализе необходимо провести $N=p^2$ (при $p=k$) опытов. Все возможные сочетания уровней и факторов дают так называемый полный факторный эксперимент (ПФЭ). Так, для двухфакторного дисперсионного анализа ДА на трех уровнях ПФЭ дает $N=3^2 = 9$ опытов, для трехфакторного ДА на трех уровнях ПФЭ дает $N=3^3 = 27$ опытов и т. д.

Возможно сокращение количества опытов и использование дробного факторного эксперимента (ДФЭ). Количество опытов в дисперсионном анализе сокращается в $1/p$ раз, если использовать в плане эксперимента латинский квадрат. Так, трехфакторный ДА по плану ДФЭ даст всего $1/3 \cdot 3^3 = 9$ опытов.

Определение. Латинским квадратом называется такая квадратная матрица $p \times p$ из p элементов, что каждый из элементов встречается в каждой строке и каждом столбце только один раз.

Из трех элементов образуется латинский квадрат 3×3 :

$$\begin{array}{ll} ABC & 123 \\ BCA \text{ или} & 231 \\ CAB & 312; \end{array} \quad (4.55)$$

$$\begin{array}{ll} ABCD & 1234 \\ BCDA \text{ или} & 2341 \\ CDAB & 3412 \\ DABC & 4123. \end{array} \quad (4.56)$$

В дисперсионном анализе используются стандартные латинские квадраты [см. формулы (4.55), (4.56)], у которых первая строка и первый столбец построены в алфавитном порядке (элементы — буквы) или в порядке натурального ряда (элементы — цифры). Способ построения таких квадратов — одношаговая циклическая перестановка.

Применение. К планированию эксперимента по схеме дисперсионного анализа с латинским квадратом прибегают обычно при исследовании влияния на процесс трех дискретно-меняющихся факторов A , B , C . При этом факторы A и B могут быть связаны с переменными, а фактор C — с помехами.

Исходные данные. Необходимо выяснить влияние трех источников изменчивости — факторов A, B, C : фактор A — строка плана, фактор B — столбец плана, фактор C — элемент плана. Результаты эксперимента можно представить моделью

$$\hat{y}_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha_i\beta_j + \alpha_i\gamma_k + \beta_j\gamma_k + \alpha_i\beta_j\gamma_k + e_{ijk}, \quad (4.57)$$

в которой имеются кроме линейных эффектов еще три парных эффекта и один тройной эффект взаимодействия факторов; e_{ijk} — ошибка с нормальным распределением. Если предположить, что взаимодействия факторов незначимы, то результаты эксперимента можно представить моделью

$$\hat{y}_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}, \quad (4.58)$$

где $\mu = \bar{y}$ — общий эффект во всех опытах; α_i — эффект фактора A (строки); β_j — эффект фактора B (столбца); γ_k — эффект фактора C (элемента квадрата).

План эксперимента без повторных опытов представлен в табл. 4.11.

Расчет итогов по строкам A_i , столбцам B_j и элементам квадрата C_k :

$$A_1 = y_1 + y_2 + y_3; A_2 = y_4 + y_5 + y_6; A_3 = y_7 + y_8 + y_9; \quad (4.59)$$

$$B_1 = y_1 + y_4 + y_7; B_2 = y_2 + y_5 + y_8; B_3 = y_3 + y_6 + y_9; \quad (4.60)$$

$$C_1 = y_1 + y_6 + y_9; C_2 = y_2 + y_4 + y_9; C_3 = y_3 + y_5 + y_7. \quad (4.61)$$

Расчет суммы квадратов всех наблюдений:

$$S_1 = \sum_{i=1}^p \sum_{j=1}^p (y_{ijk})^2 = \sum_{i=1}^{p^2} (y_{ijk})^2. \quad (4.62)$$

Расчет сумм квадратов итогов по строкам, столбцам и элементам квадрата, деленных на соответствующее число элементов:

$$S_2 = \frac{1}{p} \sum_{i=1}^p A_i^2; \quad (4.63)$$

$$S_3 = \frac{1}{p} \sum_{j=1}^p B_j^2; \quad (4.64)$$

$$S_4 = \frac{1}{p} \sum_{k=1}^p C_k^2. \quad (4.65)$$

Таблица 4.11

Трехуровневый латинский квадрат

Фактор А	Фактор В			Итоги по строкам
	b_1	b_2	b_3	
a_1	c_1 y_1	c_2 y_2	c_3 y_3	A_1
a_2	c_2 y_4	c_3 y_5	c_1 y_6	A_2
a_3	c_3 y_7	c_1 y_8	c_2 y_9	A_3
Итоги по столбцам	B_1	B_2	B_3	Общий итог
Итоги по элементам	C_1	C_2	C_3	

Расчет корректирующего члена, равного квадрату общего итога, деленное на общее число опытов:

$$S_5 = \frac{1}{n^2} G^2. \tag{4.66}$$

Расчет сумм квадратов:

$$S_A = S_2 - S_5 \text{ — для строки;} \tag{4.67}$$

$$S_B = S_3 - S_5 \text{ — для столбца;} \tag{4.68}$$

$$S_C = S_4 - S_5 \text{ — для элементов квадрата;} \tag{4.69}$$

$$S_{\text{общ}} = S_1 - S_5 \text{ — общая;} \tag{4.70}$$

$$S_{\text{ост}} = S_{\text{общ}} - (S_A + S_B + S_C) \text{ — остаточная.} \tag{4.71}$$

Расчет дисперсий:

$$s_A^2 = \frac{S_A}{p-1} \text{ — фактора } A \tag{4.72}$$

$$s_B^2 = \frac{S_B}{p-1} \text{ — фактора } B \tag{4.73}$$

$$s_C^2 = \frac{S_C}{p-1} \text{ — фактора } C \tag{4.74}$$

$$s_0^2 = \frac{S_{ост}}{(p-1)(p-2)} \quad \text{- ошибка опыта (4.75)}$$

Оценка дисперсий осуществляется по критерию Фишера. Составляются дисперсионные отношения (расчетные)

$$\frac{s_A^2}{s_0^2} = F_{pA}, \quad \frac{s_B^2}{s_0^2} = F_{pB}, \quad \frac{s_C^2}{s_0^2} = F_{pC}. \quad (4.76)$$

Принятие решения. Если расчетные значения критерия Фишера удовлетворяют неравенствам

$$F_{pA} < F_T(f_A = p-1, f_0 = (p-1)(p-2), q = 0,05); \quad (4.77)$$

$$F_{pB} < F_T(f_A, f_0, q); \quad (4.78)$$

$$F_{pC} < F_T(f_A, f_0, q), \quad (4.79)$$

то влияние всех факторов несущественно.

Если какое-нибудь из отношений (4.76) оказывается большим F_T , то влияние этого фактора существенно.

Примечание. Для четырех факторов используется дисперсионный анализ по схеме греко-латинского квадрата, при большем числе уровней — гипергреко-латинские квадраты и латинские кубы.

Пример 3. Рассмотрим задачу исследования процесса радикальной полимеризации. Решалась та же задача, что и в примере 2. Однако предполагалась оценка кроме фактора A (типа галоидного алкила на уровнях a_1, a_2, a_3, a_4) и фактора B (типа растворителя на уровнях b_1, b_2, b_3, b_4) еще и фактора C — отношения количества мономера к растворителю. Дисперсионный анализ трех факторов по схеме латинского квадрата, как известно, более экономичен по числу опытов, чем трехфакторный дисперсионный анализ. Эксперимент проводился без повторных опытов. Требовалось оценить влияние факторов A, B и C на выход полимера. Результаты эксперимента представлены в табл. 4.12.

Расчет итогов по строкам и столбцам (формулы (4.59), (4.60)) включен в табл. 4.12. Итоги по латинской букве рассчитывались по формуле (4.61):

$$c_1 = 70,5; \quad c_2 = 135,7; \quad c_3 = 116,9; \quad c_4 = 43,3.$$

Таблица 4.12.

Результаты реализации трехуровневого латинского квадрата

Уровни фактора А	Уровни фактора В						Средние по строкам
	b_1	b_2	b_3	b_4			
a_1	c_1 13,2	c_2 2,7	c_3 49,1	c_4 7,2			72,2
a_2	c_2 19,0	c_3 8,0	c_4 15,5	c_1 9,5			52,0
a_3	c_3 4,6	c_4 5,9	c_1 31,5	c_2 53,1			95,1
a_4	c_4 14,7	c_1 16,3	c_2 60,9	c_3 55,2			147,1
Средние по столбцам	51,5	32,9	157,0	125,0			Общее среднее $G = 366,4$

Расчет сумм квадратов (по формулам (4.62)—(4.66)):

$$S_1 = 13,2^2 + 2,7^2 + \dots + 55,2^2 = 14\,505,14;$$

$$S_2 = \frac{1}{4} (72,2^2 + \dots + 147,1^2) = 9649,82;$$

$$S_3 = \frac{1}{4} (51,5^2 + \dots + 125,0^2) = 11\,002,16;$$

$$S_4 = \frac{1}{4} (70,5^2 + \dots + 43,3^2) = 9731,31;$$

$$S_5 = \frac{1}{4^2} (366,4)^2 = 8390,56.$$

Расчет сумм квадратов (по формулам (4.67) —(4.71)):

$$S_A = 1259,26 \text{ для строки};$$

$$S_B = 2611,60 \text{ для столбца};$$

$$S_C = 1340,75 \text{ для элементов};$$

$$S_{\text{общ}} = 6114,58 \text{ общая};$$

$$S_{\text{ост}} = 902,97 \text{ остаточная}.$$

Расчет дисперсий (по формулам (4.72) — (4.75)):

$$s_A^2 = \frac{1}{3} \cdot 1259,26 = 419,75 \text{ фактора } A;$$

$$s_B^2 = \frac{1}{3} \cdot 2611,60 = 870,53 \text{ фактора } B;$$

$$s_C^2 = \frac{1}{3} \cdot 1340,75 = 446,92 \text{ фактора } C;$$

$$s_0^2 = \frac{1}{(4-1)(4-2)} \cdot 902,97 = 150,5 \text{ ошибка опыта.}$$

Оценка дисперсий проводится по формулам (4.76):

$$F_{pA} = s_A^2 / s_0^2 = 419,75 / 150,5 = 2,789;$$

$$F_{pB} = s_B^2 / s_0^2 = 870,0 / 150,5 = 5,78;$$

$$F_{pC} = s_C^2 / s_0^2 = 446,92 / 150,5 = 2,97.$$

Табличное значение критерия Фишера $F_T = 4,76$ ($f_1 = 3, f_0 = 6, q = 0,05$), т. е. следует признать влияние факторов A и C (тип галоидного алкила и отношение количеств мономера и растворителя) несущественным. На процесс влияет только фактор B (тип растворителя).

4.5. Регрессионный анализ

По результатам эксперимента над объектом исследования, входы и выходы которого известны, можно получить математическую модель определенного вида. Полученная *математическая модель* (если это полином определенной степени) называется *уравнением регрессии*.

Степень приближения уравнения регрессии к реальному объекту зависит не только от экспериментальных данных, но и от метода построения полинома. В качестве такого метода обычно выбирают метод наименьших квадратов, являющийся частным случаем метода максимума правдоподобия для случайных переменных с нормальным распределением. Уравнение регрессии, полученное по методу наименьших квадратов, подвергается статистическому анализу, основанному на оценках дисперсий: *проверяется однородность дисперсий, значимость коэффициентов и адекватность уравнения регрессии. Это и есть регрессионный анализ.*

В основе регрессионного анализа лежит несколько статистических предпосылок, выполнение которых гарантирует достоверность анализа полученной математической модели:

1. Выходная переменная — случайная величина с нормальным распределением; факторы — суть неслучайные величины; практически

это означает, что ошибки в управлении факторами по крайней мере на порядок меньше ошибок при измерении выходной переменной.

2. Связь между факторами отсутствует.
3. Дисперсии выходной переменной однородны (равноточны) в любой точке факторного пространства.
4. Исследуемый объект лишен динамических свойств (рассматриваются стационарные режимы объекта).

По назначению можно назвать три типа задач, для решения которых используются модели регрессионного типа.

Задача 1. Обнаружение зависимости $y = f(x)$. То есть установление самого факта наличия или отсутствия статистически значимой связи между x и y . Считаем, что между y и x существует статистически значимая взаимосвязь, если хотя бы один из коэффициентов регрессии b_i статистически значимо отличен от нуля.

Задача 2. Предсказание значений y для значений x , не вошедших в экспериментальный набор. По уравнению регрессии можно определить (предсказать) значения y для любого x , не вошедшего в набор экспериментальных точек, и определить доверительный интервал для этого значения при уровне значимости α . Часто этим пользуются для задач экстраполяции.

Задача 3. Объяснение зависимости $y = f(x)$, то есть выявление причинно-следственных связей между y и x . В задачах данного типа на основе знания коэффициентов регрессии необходимо объяснить (трактовать) на предметном уровне зависимость y от отдельных x , указать пути управления значениями y при изменении значений x . Такая постановка задачи претендует на проникновение в «физический» смысл изучаемых статистических связей. Однако в большинстве случаев статистическая модель представляет собой модель «черного ящика», которая не позволяет выявить истинные причинно-следственные связи в изучаемой системе. Это касается, в основном, моделей множественной регрессии. При этом не следует путать формальную степень влияния отдельных x на y (выражаемую парным или частным коэффициентом корреляции) и фактическую, предметную взаимосвязь x и y .

Для моделей простой линейной регрессии все три перечисленных типа задач являются фактически одной и той же. Для моделей множественной линейной регрессии задача предсказания и задача объяснения являются разными задачами.

4.5.1. Метод наименьших квадратов (МНК).

Статистики разработали много разнообразных методов обработки результатов экспериментов. Но наибольшее распространение получил метод наименьших квадратов, который был разработан более 100 лет назад Лежандром и Гауссом. Разберемся в этом методе.

Пусть имеется выборка эксперимента объемом N . Задан некоторый класс функций $f(x)$ с числом связей l , равным числу коэффициентов, которые необходимо определить. Требуется найти уравнение регрессии.

Пояснение. Наилучшее уравнение регрессии дает та функция из рассматриваемого класса, для которой сумма квадратов

$$S = \sum_{u=1}^N [y_u - f(x_u)]^2 \tag{4.80}$$

имеет наименьшее значение.

Если

$$\hat{f}(x_u) = \hat{y} = f(x_u, b_0, b_1, b_2, \dots) \tag{4.81}$$

дифференцируема и требуется найти

$$S = \sum_{u=1}^N [y_u - f(x_u, b_0, b_1, b_2, \dots)]^2 \rightarrow \min, \tag{4.82}$$

то необходимым условием минимума S является выполнение равенств

$$\frac{\partial S}{\partial b_0} = 0, \frac{\partial S}{\partial b_1} = 0, \dots \tag{4.83}$$

или

$$\left. \begin{aligned} \sum_{u=1}^N 2[y_u - f(x_u, b_0, b_1, b_2, \dots)] \frac{\partial f(x_u)}{\partial b_0} &= 0 \\ \sum_{u=1}^N 2[y_u - f(x_u, b_0, b_1, b_2, \dots)] \frac{\partial f(x_u)}{\partial b_1} &= 0 \\ \dots & \dots \end{aligned} \right\} \tag{4.84}$$

Система (4.84) состоит из уравнений, число которых равно числу неизвестных коэффициентов b_0, b_1, b_2, \dots и называется системой нормальных уравнений.

Выбор полинома $f(x_u, b_0, b_1, \dots)$ зависит от имеющейся у исследователя априорной информации и является в значительной степени искусством.

Предположим, что полином (4.81) имеет вид

$$\hat{y} = b_0 + b_1 x. \tag{4.85}$$

Это хорошо известное уравнение прямой линии.

Требуется найти коэффициенты b_0 и b_1 (в данном случае число связей $l=2$).

В целом регрессионный анализ не обязательно рассчитан на планирование эксперимента, хотя использование планирования, как будет показано в дальнейших алгоритмах, значительно расширяет его возможности. Здесь план эксперимента не используется. Выборка N получается пассивным наблюдением за объектом, хотя наблюдения и ведется по плану.

Расчет коэффициентов регрессии. Если бы все экспериментальные точки лежали строго на прямой линии, то для каждой из них было бы справедливо равенство

$$y_i - b_0 - b_1 x_{1i} = 0,$$

где $i=1, 2, \dots, N$ — номер опыта. Тогда не было бы никакой проблемы. На практике это равенство нарушается и вместо него приходится писать

$$y_i - b_0 - b_1 x_{1i} = \xi_i,$$

где ξ_i — разность между экспериментальным и вычисленным по уравнению регрессии значениями y в i -й экспериментальной точке. Эту величину иногда называют **невязкой**.

Действительно, невязка возникает по двум причинам: из-за ошибки эксперимента и из-за непригодности модели. Причем эти причины смешаны и мы не можем, не получив дополнительной информации, сказать, какая из них преобладает.

Можно *постулировать*, что модель пригодна. Тогда невязка будет порождаться только ошибкой опыта. (Еще можно, конечно, постулировать, что ошибка опыта равна нулю. Тогда невязка будет связана только с пригодностью модели, и пригодной будет такая модель, для которой все невязки равны нулю.)

Обычно оценивают независимо ошибку опыта и проверяют пригодность модели.

Мы хотим найти такие коэффициенты регрессии, при которых невязки будут минимальны. Это требование можно записать по-разному. В зависимости от этого мы будем получать разные оценки коэффициентов. Приведем одну из возможных записей

$$U = \sum_{i=1}^N \xi_i^2 = \min,$$

которая приводит к методу наименьших квадратов.

Возможен и метод наименьших кубов

$$\sum_{i=1}^N |\xi_i^2| = \min,$$

так как условие, которое мы выбираем, произвольно. Однако этот метод хуже МНК, так как мы будем получать оценки коэффициентов со значительно меньшей точностью. Да и в вычислительном отношении этот путь сложнее.

Существует и метод, в котором минимизируется сумма модулей (абсолютных величин) невязок. Но этот путь связан с дополнительными вычислительными трудностями. Условие МНК — это удачный компромисс.

Существуют другие подходы. Можно, например, минимизировать модуль максимальной невязки. Это записывается так:

$$\min \max_i |\xi_i|.$$

Предложений можно сделать довольно много, но мы не будем более на них останавливаться и перейдем непосредственно к МНК.

Когда мы ставим эксперимент, то обычно стремимся провести больше (во всяком случае не меньше) опытов, чем число неизвестных коэффициентов. Поэтому система линейных уравнений

$$\xi_i = y_i - b_0 - b_1 x_{1i}$$

оказывается переопределенной и часто противоречивой (т. е. она может иметь бесконечно много решений или может не иметь решений). **Переопределенность возникает, когда число уравнений больше числа неизвестных; противоречивость — когда некоторые из уравнений несовместимы друг с другом.**

Только если все экспериментальные точки лежат на прямой, то система становится определенной и имеет единственное решение.

МНК обладает тем замечательным свойством, что он делает **определенной любую произвольную систему уравнений**. Он делает число уравнений равным числу неизвестных коэффициентов.

Наше уравнение регрессии имеет вид

$$y = b_0 + b_1 x_1.$$

В нем два неизвестных коэффициента. Значит, применяя МНК, мы получим два уравнения.

Выше мы писали

$$U = \sum_{i=1}^N \xi_i^2 = \min.$$

Это соотношение можно записать в другом виде

$$U = \sum_{i=1}^N \xi_i^2 = \sum_{i=1}^N (y_i - b_0 - b_1 x_{1i})^2 = \min.$$

Из курса математики известно, что минимум некоторой функции, если он существует, достигается при одновременном равенстве нулю частных производных по всем неизвестным, т. е.

$$\frac{\partial U}{\partial b_0} = 0, \quad \frac{\partial U}{\partial b_1} = 0.$$

Запишем наши уравнения для определения коэффициентов. Они выглядят так :

$$-2 \sum_{i=1}^N (y_i - b_0 - b_1 x_{1i}) = 0, \quad -2 \sum_{i=1}^N (y_i - b_0 - b_1 x_{1i}) x_{1i} = 0.$$

Для вычислений раскроем скобки и проведем простые преобразования (для этого воспользуемся выражением (4.84), которые дают

$$N b_0 + \sum_{i=1}^N x_{1i} b_1 = \sum_{i=1}^N y_i, \quad \sum_{i=1}^N x_{1i} b_0 + \sum_{i=1}^N x_{1i}^2 b_1 = \sum_{i=1}^N y_i x_{1i}.$$

Окончательные формулы для вычисления коэффициентов регрессии, которые удобно находить с помощью определителей, имеют вид

$$b_0 = \frac{\sum_{i=1}^N y_i \sum_{i=1}^N x_{1i}^2 - \sum_{i=1}^N y_i x_{1i} \sum_{i=1}^N x_{1i}}{N \sum_{i=1}^N x_{1i}^2 - \left(\sum_{i=1}^N x_{1i} \right)^2};$$

$$b_1 = \frac{N \sum_{i=1}^N y_i x_{1i} - \sum_{i=1}^N y_i \sum_{i=1}^N x_{1i}}{N \sum_{i=1}^N x_{1i}^2 - \left(\sum_{i=1}^N x_{1i} \right)^2}.$$

(4.86)

Результаты эксперимента представляются следующей матрицей (табл. 4.13).

Таблица 4.13

Условия и результаты опытов

Номер опыта	x_i	y
1	x_{11}	y_1
2	x_{12}	y_2
·	·	·
·	·	·
·	·	·
i	x_{1i}	y_i
·	·	·
·	·	·
·	·	·
N	x_{1N}	y_N

Для выполнения вычислений ее расширяют так, как представлено в табл. 4.14.

Таблица 4.14

Расчетная таблица для вычислений коэффициентов регрессии

Номер опыта	x_i	y	x_i^2	yx_i	y^2	$x_i + y$	$(x_i + y)^2$
1	x_{11}	y_1	x_{11}^2	$y_1 x_{11}$	y_1^2	$x_{11} + y_1$	$(x_{11} + y_1)^2$
2	x_{12}	y_2	x_{12}^2	$y_2 x_{12}$	y_2^2	$x_{12} + y_2$	$(x_{12} + y_2)^2$
·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·
i	x_{1i}	y_i	x_{1i}^2	$y_i x_{1i}$	y_i^2	$x_{1i} + y_i$	$(x_{1i} + y_i)^2$
·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·
N	x_{1N}	y_N	x_{1N}^2	$y_N x_{1N}$	y_N^2	$x_{1N} + y_N$	$(x_{1N} + y_N)^2$
Σ	$\sum_{i=1}^N x_{1i}$	$\sum_{i=1}^N y_i$	$\sum_{i=1}^N x_{1i}^2$	$\sum_{i=1}^N y_i x_{1i}$	$\sum_{i=1}^N y_i^2$	—	$\sum_{i=1}^N (x_{1i} + y_i)^2$
Среднее значение	\bar{x}_1	\bar{y}					

Как видно, в этой таблице сделано больше вычислений, чем требуется для расчета b_0 и b_1 . Эти «лишние» данные нужны для проверки правильности расчетов.

Возможны два способа проверки. Первый из условия

$$\sum_{i=1}^N (x_{1i} + y_i)^2 = \sum_{i=1}^N x_{1i}^2 + 2 \sum_{i=1}^N y_i x_{1i} + \sum_{i=1}^N y_i^2,$$

которое хорошо известно из школьной математики. (Оно должно выполняться не только для сумм, но и в каждой строчке таблицы.) Второй способ использует условие $\bar{y} = b_0 + b_1 \bar{x}_1$. Подставляя в это соотношение \bar{y} и \bar{x}_1 из последней строки таблицы и один из коэффициентов, можно найти другой коэффициент и сравнить с расчетным.

Вторая из проверок является наиболее полной, наиболее жесткой. Она проверяет не только вычисления сумм, но и вычисления коэффициентов

На практике используют обе проверки, чтобы в случае ошибок в таблице не считать зря коэффициенты.

Необходимо помнить: *никакая проверка не гарантирует вас от ошибок в записи исходных данных.*

Следует иметь в виду: *никакие результаты вычислений нельзя ни использовать, ни обсуждать, пока они не проверены.* Иначе можно впасть в заблуждение и, в лучшем случае, потерять время.

Нанесем исходные данные и полученное уравнение на график (рис. 4.8).

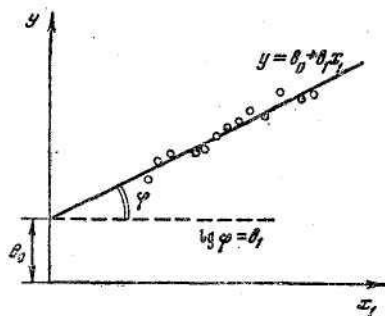


Рис. 4.8. Линейное уравнение регрессии

Выделим для удобства рассмотрения несколько экспериментальных точек и отрезок нашего уравнения в большем масштабе (рис. 4.9).

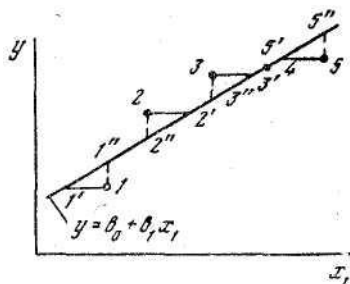


Рис. 4.9 Линейное уравнение регрессии (фрагмент)

Мы выбрали пять экспериментальных точек, которые пронумеровали цифрами 1, 2, 3, 4, 5. Четвертая точка оказалась лежащей на линии. МНК состоит в том, чтобы минимизировать сумму квадратов отрезков, характеризующих расхождение между экспериментальными точками и полученным уравнением. Мы минимизировали сумму квадратов пунктирных отрезков.

Если бы наше уравнение регрессии имело вид

$$x_1 = b_0 + b_1 y,$$

то мы минимизировали бы сумму сплошных отрезков.

Во всех формулах тогда пришлось бы x_1 и y поменять местами и коэффициенты получились бы другими (если, конечно, не все невязки равны нулю).

Мы находим невязки по оси y , поэтому и минимизируется сумма квадратов вертикальных отрезков. Обе линии совпадут только в том случае, если все невязки равны нулю, т. е. если все экспериментальные точки лежат точно на прямой линии.

Теперь мы можем узнать, какая же получилась сумма квадратов невязок. Будем называть ее **остаточной суммой квадратов**.

Из рисунков видно, что для этого надо вычислить по уравнению значения y в условиях каждого опыта. Будем называть такое значение **предсказанным** и обозначать y . Затем надо найти все невязки (отрезки), возвести их в квадрат и сложить (табл. 4.15).

Таблица 4.15

Расчет остаточной суммы квадратов

Номер опыта	y	\hat{y}	$\Delta y = y - \hat{y}$	Δy^2
1	y_1	\hat{y}_1	Δy_1	Δy_1^2
2	y_2	\hat{y}_2	Δy_2	Δy_2^2
⋮	⋮	⋮	⋮	⋮
i	y_i	\hat{y}_i	Δy_i	Δy_i^2
⋮	⋮	⋮	⋮	⋮
N	y_N	\hat{y}_N	Δy_N	Δy_N^2
				$\sum_{i=1}^N \Delta y_i^2$

Величина $\sum_{i=1}^N \Delta y_i^2$ и есть остаточная сумма квадратов, которую

мы раньше обозначили $\sum_{i=1}^N \epsilon_i^2$. МНК гарантирует, что эта величина минимально возможная.

Итак, мы научились находить наилучшие в смысле МНК оценки коэффициентов линейного уравнения для одного фактора. Обобщение на многофакторный случай не связано с какими-либо принципиальными трудностями. При этом вычисления значительно усложняются и требуют привлечения аппарата алгебры матриц, фрагменты которой будут рассмотрены в дальнейшем. А пока мы воспользуемся тем, что наши матрицы планирования ортогональны. Далее будем рассматривать только этот случай, который позволяет резко упростить вычисления, что составляет одно из преимуществ планирования эксперимента. Можно показать, что для любого числа факторов коэффициенты будут вычисляться по формуле

$$b_j = \frac{\sum_{i=1}^N y_i x_{ji}}{N}$$

В этой формуле $j=0,1,2,\dots,k$ — номер фактора. Ноль записан для вычисления b_0 . Воспользуемся формулами (4.86) для вычисления коэффициентов регрессии. В первой формуле

$$\sum_{i=1}^N x_{li} = 0$$

в силу симметричности плана. Поэтому после сокращения формулы приобретают вид

$$b_0 = \frac{\sum_{i=1}^N y_i}{N}, \quad b_1 = \frac{\sum_{i=1}^N y_i x_{1i}}{\sum_{i=1}^N x_{1i}^2},$$

где

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

что совпадает с написанным выше.

Так как каждый фактор (кроме x_0) варьируется на двух уровнях + 1 и -1, то вычисления сводятся к приписыванию столбцу y знаков соответствующего фактору столбца и алгебраическому сложению полученных значений. Деление результата на число опытов в матрице планирования дает искомый коэффициент. Это очень простая формула, но ею необходимо научиться пользоваться безошибочно.

4.5.2. Статистический анализ уравнения регрессии.

В случае получения уравнения регрессии (4.85) с коэффициентами (4.86) необходимо провести статистический анализ коэффициентов и уравнения в целом.

Пояснение. Статистический анализ уравнения регрессии включает три этапа: проверку однородности дисперсий в каждой точке экспериментальной выборки N ; оценку значимости всех коэффициентов регрессии; проверку адекватности уравнения регрессии экспериментальным данным.

Учитывая, что при пассивном эксперименте наблюдения дублировать принципиально невозможно, проверку однородности дисперсии провести нельзя. Таким образом, нарушается одна из предпосылок регрессионного анализа, что влечет за собой ряд трудностей. В частности, невозможно точно оценить значимость коэффициентов и адекватность уравнения. Однако иногда удается получить два-три наблюдения для идентичных значений факторов (в одной точке факторного пространства). И хотя это не дает возможности проверить однородность дисперсий, но позволяет сделать другие оценки.

Оценка значимости коэффициентов регрессии производится двумя равноценными способами: по критерию Стьюдента или построением доверительного интервала.

При оценке значимости коэффициентов регрессии по критерию Стьюдента сначала необходимо рассчитать дисперсию воспроизводимости (ошибку опыта):

$$s_0^2 = \frac{1}{N_0 - 1} \sum_{k=1}^{N_0} (y_{0k} - \bar{y}_0)^2,$$

где y_{0k} — значение переменной в точке факторного пространства; N_0 — число параллельных опытов в этой точке;

$$\bar{y}_0 = \frac{1}{N_0} \sum_{k=1}^{N_0} y_{0k} \quad \text{— среднее значение переменной в этой точке;}$$

среднеквадратические отклонения коэффициентов регрессии

$$s_{b_0} = \sqrt{\frac{s_0^2 \sum_{u=1}^N x_u^2}{N \sum_{u=1}^N x_u^2 - \left(\sum_{u=1}^N x_u \right)^2}};$$

$$s_{b_1} = \sqrt{\frac{s_0^2 N}{N \sum_{u=1}^N x_u^2 - \left(\sum_{u=1}^N x_u \right)^2}}.$$

Расчет t -критерия Стьюдента:

$$t_{ip} = \frac{|b_i|}{s_{b_i}},$$

где b_i — i -ый коэффициент регрессии.

Принятие решений. Если окажется, что

$$t_{ip} > t_{\tau}(f_0 = N_0 - 1, q = 0,05)$$

для числа степеней свободы f_0 и заданного уровня значимости q , то i -ый коэффициент признается значимым. Незначимые коэффициенты могут исключаться из уравнения регрессии. Оставшиеся коэффициенты пересчитываются заново.

При оценке значимости коэффициентов регрессии по методу построения доверительного интервала сначала необходимо найти дисперсию коэффициента регрессии $s_{\{b_j\}}^2$. Она определяется в нашем случае по формуле

$$s_{\{b_j\}}^2 = \frac{s_{\{y\}}^2}{N},$$

если параллельные опыты отсутствуют.

Из формулы видно, что дисперсии всех коэффициентов равны друг другу, так как они зависят только от ошибки опыта и числа опытов.

Теперь легко построить доверительный интервал (Δb_j)

$$\Delta b_j = \pm ts_{(b_j)}.$$

Здесь t — табличное значение критерия Стьюдента при числе степеней свободы, с которыми определялась $s^2_{(y)}$, и выбранном уровне значимости (обычно 0,05); $s_{(b_j)}$ — квадратичная ошибка коэффициента регрессии

$$s_{(b_j)} = \pm \sqrt{s^2_{(b_j)}}.$$

Формулу для доверительного интервала можно записать в следующей эквивалентной форме:

$$\Delta b_j = \pm \frac{ts_{(y)}}{\sqrt{N}}.$$

Коэффициент значим, если его абсолютная величина больше доверительного интервала. Доверительный интервал задается верхней и нижней границами $b_j + \Delta b_j$ и $b_j - \Delta b_j$.

Для отыскания значений t -критерия можно воспользоваться таблицей, фрагмент из которой приведен в табл. 4.16.

Таблица 4.16

Значения t -критерия Стьюдента при 5%-ном уровне значимости

Число степеней свободы	Значения t -критерия	Число степеней свободы	Значения t -критерия	Число степеней свободы	Значения t -критерия
1	12,71	11	2,201	21	2,080
2	4,303	12	2,179	22	2,074
3	3,182	13	2,160	23	2,069
4	2,776	14	2,145	24	2,064
5	2,571	15	2,131	25	2,060
6	2,447	16	2,120	26	2,056
7	2,365	17	2,110	27	2,052
8	2,306	18	2,101	28	2,048
9	2,262	19	2,093	29	2,045
10	2,228	20	2,086	30	2,042
				∞	1,960

Таблица построена следующим образом. Столбцы соответствуют различным степеням свободы и значениям критерия.

Пусть в двух разных задачах случайно оказались два численно равных коэффициента регрессии. Доверительные интервалы для них оказались различными. Из них значим только второй

Задача	b_j	Δb_j
1	5,3	$\pm 5,5$
2	5,3	$\pm 2,6$

В действительности чем уже доверительный интервал (при заданном α), тем с большей уверенностью можно говорить о значимости коэффициента.

Есть рабочее правило: если абсолютная величина коэффициента больше, чем доверительный интервал, то коэффициент значим.

Возможно проверять значимость коэффициентов по t -критерию. Для этого воспользуемся формулой

$$t = \frac{|b_j|}{s\{b_j\}}$$

Вычисленное значение t -критерия сравнивается с табличным при заданном α и соответствующем числе степеней свободы. Полученные выводы о значимости коэффициентов должны совпадать с предыдущими.

Так производится проверка значимости коэффициентов.

Проверка адекватности уравнения регрессии. Для проверки необходимо рассчитать остаточную сумму квадратов

$$S_{\text{ост}} = \sum_{u=1}^N (y_u - \hat{y}_u)^2,$$

где y_u — экспериментальные значения выходной переменной;

\hat{y}_u — значения, рассчитанные по уравнению регрессии.

Число степеней свободы

$$f_{\text{ад}} = N - l,$$

где l — число связей, равное числу коэффициентов уравнения, оставшихся после проверки их значимости. Тогда дисперсия адекватности

$$s_{\text{ад}}^2 = \frac{S_{\text{ост}}}{f_{\text{ад}}};$$

адекватность уравнения проверяется по критерию Фишера:

$$F_p = \frac{s_{\text{ад}}^2}{s_0^2}.$$

Принятие решений. Если расчетное значение критерия Фишера

$$F_p < F_T(f_{\text{ад}} = N - l, f_0 = N_0 - 1, q = 0,05) \quad (4.87)$$

для степеней свободы $f_{\text{ад}}$, f_0 и заданного уровня значимости, то уравнение считается адекватным. Если условие (4.87) не выполняется,

то уравнение считается неадекватным и требуется изменение полинома $f(x_i)$ — вводятся новые факторы, изменяется порядок полинома.

Рассмотрим вопрос о проверке адекватности модели на примере

На рис.4.10 а,б приведены два рисунка с одинаковым расположением экспериментальных точек и, следовательно, одинаковым разбросом относительно линии регрессии но с различным средним разбросом в точках (с различной дисперсией воспроизводимости).

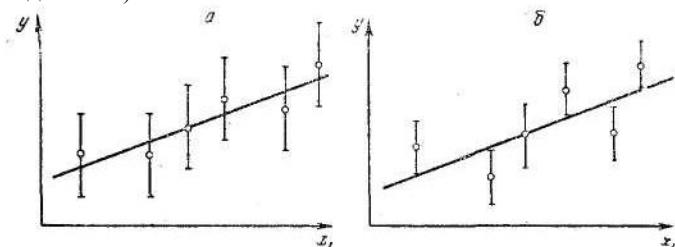


Рис. 4.10. Проверка адекватности

Разброс в точках показан, как это иногда делается, отрезками прямых, составляющих доверительный интервал, равный $+2s_{(y)}$. Модель можно считать адекватной только в первом случае.

В данном случае разброс в точках такого же порядка, что и разброс относительно линии. Поэтому можно предполагать, что построенная модель пригодна. (Дальше мы выясним, как проверить это количественно.) Во втором случае опыты «слишком» точны. Требуется более сложная модель, чтобы точность ее предсказания была сравнима с точностью эксперимента.

Это качественные соображения, а нам нужна количественная мера.

Для характеристики среднего разброса относительно линии регрессии вполне подходит остаточная сумма квадратов. Неудобство состоит в том, что она зависит от числа коэффициентов в уравнении: если ввести столько коэффициентов, сколько проведено независимых опытов, то получим остаточную сумму, равную нулю. Поэтому предпочитают относить ее на один «свободный» опыт. **Число таких опытов называется числом степеней свободы (f).**

Числом степеней свободы в статистике называется разность между числом опытов и числом коэффициентов (констант), которые уже вычислены по результатам этих опытов независимо друг от друга.

Если, например, вы провели факторный эксперимент 2^3 и нашли линейное уравнение регрессии, то число степеней свободы

$$f = N - (k + 1) = 8 - (3 + 1) = 4.$$

Остаточная сумма квадратов, деленная на число степеней свободы, называется *остаточной дисперсией*, или *дисперсией адекватности* ($s_{\text{ад}}^2$).

$$s_{\text{ад}}^2 = \frac{\sum_{i=1}^N \Delta y_i^2}{f}.$$

Рассмотрим пример. Требуется найти число степеней свободы для $s_{\text{ад}}^2$ в следующем случае: план 2^{4-1} и четыре параллельных опыта в нулевой точке для вычисления ошибки опыта; модель линейная.

Один из возможных ответов — семь степеней свободы — основан на следующих предположениях. Прделано 12 опытов: $2^{4-1}=8$ плюс 4 нулевых. В уравнение входит 5 коэффициентов. Следовательно, $f=12-5=7$. Здесь не учтено, что параллельные опыты нельзя считать самостоятельными, так как они дублируют друг друга. Поэтому они все дают одну степень свободы. Другой неправильный ответ — 4 степени свободы. Этот неправильный ответ получится в результате следующего рассуждения. Прделано 12 опытов: восемь по матрице планирования и четыре нулевых. Так как все нулевые опыты тождественны, то они дают одну степень свободы. Число коэффициентов в модели равно пяти. Следовательно, $f=9-5=4$. Здесь не было обращено внимание на то, что опыты в нулевой точке не используются при вычислении коэффициентов и не могут поэтому входить в число степеней свободы.

Правильный ответ — три степени свободы. Действительно, мы провели 12 опытов, но четыре опыта в нулевой точке были проведены для других целей и в вычислении коэффициентов не участвовали, поэтому они не входят в число степеней свободы. (А если бы входили — такие случаи возможны, то давали бы не четыре, а только одну степень свободы.) Число коэффициентов модели — пять. Следовательно,

$$f=8-5=3$$

Следует запомнить правило: *в планировании эксперимента число степеней свободы для дисперсии адекватности равно числу различных опытов, результаты которых используются при подсчете коэффициентов регрессии, минус число определяемых коэффициентов.*

Для проверки гипотезы об адекватности можно использовать F -критерий (этот критерий уже использовался для сравнения двух дисперсий)

$$F = \frac{s_{ax}^2}{s_{(y)}^2}$$

Величину, стоящую в числителе этой формулы, мы только что научились считать, а знаменатель — это дисперсия воспроизводимости со своим числом степеней свободы. Удобство использования критерия Фишера состоит в том, что проверку гипотезы можно свести к сравнению с табличным значением. Фрагмент соответствующей таблицы, который может удовлетворить ваши нужды не только в упражнениях этой книги, но и в большинстве случаев практики, приведен ниже (табл. 4.18).

Таблица 4.18

Значения F-критерия Фишера при 5%-ном уровне значимости

f_1	$f_2=1$	2	3	4	5	6	12	24	∞
1	164,4	199,5	215,7	224,6	230,2	234,0	244,9	249,0	254,3
2	18,5	19,2	19,2	19,3	19,3	19,3	19,4	19,4	19,5
3	10,1	9,6	9,3	9,1	9,0	8,9	8,7	8,6	8,5
4	7,7	6,9	6,6	6,4	6,3	6,2	5,9	5,8	5,6
5	6,6	5,8	5,4	5,2	5,1	5,0	4,7	4,5	4,4
6	6,0	5,1	4,8	4,5	4,4	4,3	4,0	3,8	3,7
7	5,5	4,7	4,4	4,1	4,0	3,9	3,6	3,4	3,2
8	5,3	4,5	4,1	3,8	3,7	3,6	3,3	3,1	2,9
9	5,1	4,3	3,9	3,6	3,5	3,4	3,1	2,9	2,7
10	5,0	4,1	3,7	3,5	3,3	3,2	2,9	2,7	2,5
11	4,8	4,0	3,6	3,4	3,2	3,1	2,8	2,6	2,4
12	4,8	3,9	3,5	3,3	3,1	3,0	2,7	2,5	2,3
13	4,7	3,8	3,4	3,2	3,0	2,9	2,6	2,4	2,2
14	4,6	3,7	3,3	3,1	3,0	2,9	2,5	2,3	2,1
15	4,5	3,7	3,3	3,1	2,9	2,8	2,5	2,3	2,1
16	4,5	3,6	3,2	3,0	2,9	2,7	2,4	2,2	2,0
17	4,5	3,6	3,2	3,0	2,8	2,7	2,4	2,2	2,0
18	4,4	3,6	3,2	2,9	2,8	2,7	2,3	2,1	1,9
19	4,4	3,5	3,1	2,9	2,7	2,6	2,3	2,1	1,9
20	4,4	3,5	3,1	2,9	2,7	2,6	2,3	2,1	1,9
22	4,3	3,4	3,1	2,8	2,7	2,6	2,2	2,0	1,8
24	4,3	3,4	3,0	2,8	2,6	2,5	2,2	2,0	1,7
26	4,2	3,4	3,0	2,7	2,6	2,5	2,2	2,0	1,7
28	4,2	3,3	3,0	2,7	2,6	2,4	2,1	1,9	1,7
30	4,2	3,3	2,9	2,7	2,5	2,4	2,1	1,9	1,6
40	4,1	3,2	2,9	2,6	2,5	2,3	2,0	1,8	1,5
60	4,0	3,2	2,8	2,5	2,4	2,3	1,9	1,7	1,4
120	3,9	3,1	2,7	2,5	2,3	2,2	1,8	1,6	1,3
∞	3,8	3,0	2,6	2,4	2,2	2,1	1,8	1,5	1,0

Таблица построена следующим образом. Столбцы связаны с определенным числом степеней свободы для числителя f_1 , строки — для знаменателя f_2 . На пересечении соответствующих строки и столбца

стоят критические значения F-критерия. Как правило, в технических задачах используется уровень значимости 0,05.

Если рассчитанное значение F-критерия не превышает табличного, то с соответствующей доверительной вероятностью модель можно считать адекватной. При превышении табличного значения эту приятную гипотезу приходится отвергать.

Если модель адекватна, то мы можем перейти к крутому восхождению. Если нет — приходится преодолевать дополнительные трудности

Итак, в этом разделе мы освоили основные методы обработки экспериментальных данных, полученных при планировании эксперимента. Мы научились не только вычислять коэффициенты регрессии, но и проводить статистические оценки адекватности и значимости.

Мы подробно рассмотрели метод наименьших квадратов — эффективный и простой способ получения оценок коэффициентов регрессии. Эти оценки приводят к минимально возможной остаточной сумме квадратов и в этом смысле являются оптимальными.

Одновременно мы установили важное требование обязательной проверки правильности вычислений и научились выполнять это требование при применении метода наименьших квадратов.

Мы узнали, что МНК становится частью регрессионного анализа при проверке статистических гипотез. При этом должны выполняться следующие постулаты:

- 1) параметр оптимизации — случайная величина с нормальным законом распределения;
- 2) дисперсия параметра оптимизации не зависит от значений параметра оптимизации;
- 3) значения факторов — неслучайные величины;
- 4) факторы не коррелированы.

Мы выяснили, как можно проверить выполнимость этих постулатов и к чему приводит их нарушение.

Всякая модель ценна постольку, поскольку она верно отражает описываемое явление. Мы выбрали подходящий статистический метод проверки адекватности модели, основанный на критерии Фишера, и научились им пользоваться.

Кроме проверки адекватности следует проводить проверку значимости коэффициентов. Эта проверка осуществляется с помощью критерия Стьюдента. Мы рассмотрели два варианта такой проверки: с помощью построения доверительных интервалов и непосредственно сравнением с табличным значением критерия.

4.5.3. Матричный подход к регрессионному анализу

Существует множество причин, по которым целесообразно ввести язык матричной алгебры. Вот некоторые из них: многие сложные случаи, возникающие в планировании эксперимента, такие как планы с неравномерным дублированием опытов, планы с разным числом уровней и т. п., требуют матричной формы описания; рассмотрение некоторых критериев оптимальности планов естественно вести на матричном языке; почти все научные работы используют матричные обозначения. Алгебра матриц, как известно, является универсальным языком современной науки и техники. В настоящем разделе мы будем излагать только те определения и операции с матрицами, которые необходимы для рассмотрения регрессионного анализа. Более полное изложение матричной алгебры можно найти в специальной литературе.

4.5.3.1. Метод наименьших квадратов для одного фактора

Будем действовать в такой последовательности. Сначала рассмотрим численный пример линейного уравнения для одного фактора, который в общем виде рассмотрен в предыдущем разделе. Затем опишем его на матричном языке. Это позволит нам в дальнейшем сделать некоторые обобщения.

Таблица 4.19

Условия и результаты опытов

Номер опыта	x_1	y	Номер опыта	x_1	y
1	-2	0	4	+1	3
2	-1	1	5	+2	4
3	0	2			

Пусть известно, что y связан с x_1 линейным уравнением $y=b_0+b_1x_1$.

В результате проведенного эксперимента была получена следующая таблица, аналогичная табл. 4.13, в которой x_1 принимает пять кодированных значений (табл. 4.19). Произведем вычисления в соответствии с предписанием табл. 4.14 (табл. 4.20).

Таблица 4.20

Расчетная таблица вычисления коэффициентов регрессии

Номер опыта	x_i	y	x_i^2	yx_i	y^2	$x_i + y$	$(x_i + y)^2$
1	-2	0	4	0	0	-2	4
2	-1	1	1	-1	1	0	0
3	0	2	0	0	4	2	4
4	+1	3	1	3	9	4	16
5	+2	4	4	8	16	6	36
Σ	0	10	10	10	30	—	60
Среднее значение	0	.2					

Воспользовавшись полученными ранее формулами, вычислим коэффициенты:

$$b_0 = \frac{\sum_{i=1}^5 y_i \sum_{i=1}^5 x_{1i}^2 - \sum_{i=1}^5 y_i x_{1i} \sum_{i=1}^5 x_{1i}}{5 \sum_{i=1}^5 x_{1i}^2 - \left(\sum_{i=1}^5 x_{1i} \right)^2} = \frac{10 \cdot 10 - 10 \cdot 0}{5 \cdot 10 - 0} = 2;$$

$$b_1 = \frac{5 \sum_{i=1}^5 y_i x_{1i} - \sum_{i=1}^5 y_i \sum_{i=1}^5 x_{1i}}{5 \sum_{i=1}^5 x_{1i}^2 - \left(\sum_{i=1}^5 x_{1i} \right)^2} = \frac{5 \cdot 10 - 10 \cdot 0}{5 \cdot 10 - 0} = 1.$$

Теперь можно написать уравнение $y=2+1 \cdot x_1$. С помощью этого примера мы воспроизвели процедуру МНК, изложенную в предыдущем разделе.

Будем использовать теперь другой язык.

В нашем примере участвуют три множества элементов: элементы, задающие условия проведения опытов, элементы, характеризующие их результаты, и неизвестные коэффициенты, которые нужно определить.

Так, элементы, характеризующие результаты опытов, можно представить в виде столбца

$$Y = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix},$$

неизвестным коэффициентам соответствует столбец

$$\mathbf{V} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix},$$

а элементы, задающие условия опытов, удобно представить в виде таблицы

$$\mathbf{X} = \begin{array}{cc} & x_0 & x_1 \\ \begin{array}{c} + \\ + \\ + \\ + \\ + \end{array} & \begin{bmatrix} +1 & -2 \\ +1 & -1 \\ +1 & 0 \\ +1 & +1 \\ +1 & +2 \end{bmatrix} \end{array}.$$

В этой таблице появился столбец x_0 , состоящий из +1. Он введен для удобства вычислений всех коэффициентов, включая свободный член b_0 . Фактически это значит, что мы переписали исходное уравнение так:

$$y = b_0 x_0 + b_1 x_1.$$

Об этом уже говорилось ранее.

Таблицы, в которых собраны упорядоченные некоторым образом элементы, называются матрицами. Следовательно, в нашем случае \mathbf{Y} , \mathbf{V} и \mathbf{X} являются матрицами. Элементы \mathbf{Y} упорядочены по номерам опытов, элементы \mathbf{X} , кроме того, — по номерам переменных (x_0, x_1), а элементы \mathbf{V} — по номерам коэффициентов, которые соответствуют номерам переменных.

Для того чтобы воспользоваться матрицами при описании регрессионного анализа, необходимо ввести некоторые операции над ними, что мы и сделаем ниже.

4.5.3.2. Некоторые операции над матрицами

Матрицы, описанные в предыдущем параграфе, различаются по числу элементов, числу строк и числу столбцов. Так, в матрице \mathbf{Y} один столбец, пять строк и пять элементов, а в матрице \mathbf{X} тоже пять строк, но два столбца и десять элементов. Если число строк и число столбцов различны, то матрицы называются *прямоугольными*, а при равном числе строк и столбцов — *квадратными*. Все матрицы из этого примера — прямоугольные. Если матрица имеет всего один столбец, то ее называют матрицей-столбцом или вектор-столбцом. Примерами служат матрицы \mathbf{Y} и \mathbf{V} . Аналогично можно определить и матрицы-строки (векторы-строки).

Вернемся к нашему примеру. На основании исходных данных можно записать систему из пяти уравнений по одному уравнению для

каждого опыта:

$$y_i = b_0 x_0 + b_1 x_{1i}, \quad (i=1, 2, \dots, 5)$$

или в развернутой форме:

$$0 = b_0 \cdot 1 + b_1 \cdot (-2);$$

$$1 = b_0 \cdot 1 + b_1 \cdot (-1);$$

$$2 = b_0 \cdot 1 + b_1 \cdot (0);$$

$$3 = b_0 \cdot 1 + b_1 \cdot (+1);$$

$$4 = b_0 \cdot 1 + b_1 \cdot (+2).$$

На матричном языке эта система уравнений выглядит следующим образом:

$$Y = XB, \text{ т. е. } \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} +1 & -2 \\ +1 & -1 \\ +1 & 0 \\ +1 & +1 \\ +1 & +2 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Чтобы эти две записи стали эквивалентными, необходимо ввести определенные правила перемножения матриц. Будем в произведении различать матрицу, стоящую слева, и матрицу, стоящую справа. Перемножить две матрицы это значит получить матрицу произведений, элементы которой находятся по следующим правилам.

Элементы первой строки матрицы, стоящей слева, умножаются на соответствующие элементы матрицы, стоящей справа, и полученные произведения складываются. В нашем случае имеем: $(+1) b_0 + (-1) b_1$. Для получения элемента, стоящего на пересечении первого столбца и второй строки, аналогичная операция продельвается со второй строкой матрицы, стоящей слева, и тем же самым первым столбцом матрицы, стоящей справа, т. е. $(+1) b_0 + (-1) b_1$. Продолжая таким образом до последней строки матрицы, стоящей слева, получаем все элементы первого столбца матрицы произведений. Эта процедура повторяется столько раз, сколько вектор-столбцов содержит матрица, стоящая справа. В нашем случае эта матрица имеет только один столбец. Из определения видно, что матрица произведений имеет столько столбцов, сколько матрица, стоящая справа, и столько строк, сколько матрица, стоящая слева. В рассматриваемом примере матрица-произведение имеет один столбец и пять строк, что соответствует размерности матрицы Y . И тогда матрица-произведение имеет вид

$$\begin{bmatrix} 1 & b_0 + b_1(-2) \\ 1 & b_0 + b_1(-1) \\ 1 & b_0 + b_1(0) \\ 1 & b_0 + b_1(+1) \\ 1 & b_0 + b_1(+2) \end{bmatrix}.$$

Сопоставление матрицы-произведения с системой уравнений убеждает нас в тождественности матричной и нематричной форм записей. Вектор \mathbf{Y} и есть матрица произведений в данном случае. Элементы матрицы-произведения называются скалярными произведениями вектор-строки матрицы, стоящей слева, и соответствующего вектор-столбца матрицы, стоящей справа. В правилах перемножения матриц существуют особенности, не имеющие аналога в числах. Так, **небезразлично, в каком порядке записаны матрицы в произведении**. Как видно, левая и правая матрицы неравноправны. Если мы захотим умножить матрицу \mathbf{B} на матрицу \mathbf{X} (\mathbf{BX}), то убедимся, что этого сделать невозможно, ибо длины векторов, входящих в скалярное произведение, должны быть согласованы.

Таким образом, **для двух произвольных матриц произведение существует, если число столбцов матрицы, стоящей слева, равно числу строк матрицы, стоящей справа**. Ясно, что для двух квадратных матриц одинакового размера существуют оба произведения (справа и слева), однако они могут быть различными. Матрицы, произведение которых не зависит от порядка сомножителей, называются **коммутирующими**. В общем же случае для произведения матриц коммутативный закон не выполняется.

Перейдем теперь к системе нормальных уравнений МНК, которая в нашем случае выглядит следующим образом :

$$5b_0 + 0b_1 = 10; \quad 0b_0 + 10b_1 = 10.$$

Можно показать, что в матричном виде она запишется следующим образом: $\mathbf{X}^T \mathbf{X} \mathbf{B} = \mathbf{X}^T \mathbf{Y}$. Здесь \mathbf{X}^T обозначает матрицу, **транспонированную** по отношению к матрице \mathbf{X} .

Протранспонировать матрицу — это значит столбцы исходной матрицы сделать строками транспонированной матрицы, сохранив их последовательность.

Так, в нашем случае транспонированная матрица

$$\mathbf{X}^T = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 \\ -2 & -1 & 0 & +1 & +2 \end{bmatrix}.$$

Для получения системы нормальных уравнений следует умножить обе части исходной системы уравнений слева на \mathbf{X}^T .

Выполним эти операции

$$\mathbf{X}^T \mathbf{Y} = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 \\ -2 & -1 & 0 & +1 & +2 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & +1 & +2 & +3 & +4 \\ 0 & +(-1) & +0 & +3 & +8 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \end{bmatrix};$$

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 \\ -2 & -1 & 0 & +1 & +2 \end{bmatrix} \cdot \begin{bmatrix} +1 & -2 \\ +1 & -1 \\ +1 & 0 \\ +1 & +1 \\ +1 & +2 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & +1 & +1 & +1 & +1 & -2 & -1 & 0 & +1 & +2 \\ -2 & -1 & 0 & +1 & +2 & +4 & +1 & 0 & +1 & +4 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}.$$

Теперь можно записать систему уравнений:

$$\begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \end{bmatrix}.$$

Матрица $\mathbf{X}^T \mathbf{X}$ называется *матрицей системы нормальных уравнений*. Она обладает рядом важных свойств. Прежде всего заметим, что в этой матрице два элемента, расположенных симметрично относительно диагонали, идущей с левого верхнего угла в правый нижний (так называемой *главной диагонали*), равны между собой. В нашем случае это нули. Такое свойство характерно для матриц систем нормальных уравнений МНК, так как векторы, входящие в скалярные произведения, коммутативны.

Матрица, элементы которой симметричны относительно главной диагонали, называется *симметричной*. Если все элементы вне главной диагонали равны нулю, то такая матрица называется *диагональной*. В дальнейшем нам понадобится еще одна разновидность диагональных матриц. Диагональная матрица, все элементы главной диагонали которой являются единицами, называется *единичной* матрицей. Единичная матрица играет в алгебре матриц такую же роль, какую единица — в алгебре чисел.

Решить систему нормальных уравнений это значит записать в явном виде элементы вектора \mathbf{B} (b_0 и b_1). Если бы мы имели дело с числами, то для этого нужно было бы поделить обе части на коэффициент при

неизвестном и получить ответ. Но для матриц вместо деления (которое не определено) используется специальная операция умножения на *обратную* матрицу. Задача состоит в том, чтобы превратить матрицу, стоящую перед матрицей неизвестных коэффициентов, в единичную. Тогда умножение вектора **В** на единичную матрицу его не изменит, а чтобы равенство не нарушилось, и правую часть придется домножить на соответствующую матрицу. Если условиться обозначать обратную матрицу степенью -1 , то предыдущие рассуждения приведут к следующей записи:

$$(\mathbf{X}^T\mathbf{X})^{-1} (\mathbf{X}^T\mathbf{X})\mathbf{B}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Здесь система нормальных уравнений МНК умножена слева на матрицу, обратную к матрице системы нормальных уравнений. Произведение обратной матрицы на прямую справа равно единичной матрице, которую условимся обозначать **E**:

$$\mathbf{E}=(\mathbf{X}^T\mathbf{X})^{-1} (\mathbf{X}^T\mathbf{X}).$$

В этом равенстве участвуют три матрицы. Матрица системы нормальных уравнений $\mathbf{X}^T\mathbf{X}$, которую называют *прямой* матрицей, $(\mathbf{X}^T\mathbf{X})^{-1}$, — обратная матрица. Перепишем это равенство для нашей задачи

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}.$$

Неизвестные элементы обратной матрицы обозначены a_{ij} , где $i=1, 2$ соответствует строке, а $j = 1, 2$ — столбцу. Найдём эти элементы

$$\begin{aligned} 1 &= a_{11} \cdot 5 + a_{12} \cdot 0; & 0 &= a_{21} \cdot 5 + a_{22} \cdot 0; \\ 0 &= a_{11} \cdot 0 + a_{12} \cdot 10; & 1 &= a_{21} \cdot 0 + a_{22} \cdot 10. \end{aligned}$$

Отсюда следует, что

$$a_{11}=1/5, \quad a_{12}=0, \quad a_{21}=0, \quad a_{22}=1/10.$$

Запишем обратную матрицу

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 1/5 & 0 \\ 0 & 1/10 \end{bmatrix}.$$

Столь элементарной процедурой можно только в простых случаях (как наш пример). В общем случае приходится прибегать к более сложным алгоритмам и вычислительной технике.

Отметим некоторые существенные свойства обратной матрицы. Произведение прямой и обратной матрицы коммутативно. Если прямую матрицу обозначить **A**, то $\mathbf{A}\mathbf{A}^{-1}=\mathbf{A}^{-1}\mathbf{A}=\mathbf{E}$.

Матрица, обратная к симметричной, тоже будет симметрична. На главной диагонали матрицы, обратной к диагональной, будут стоять числа, обратные соответствующим числам, стоящим на диагонали прямой матрицы. Зная это свойство, мы могли не проделывать предыдущие вычисления, а сразу записать обратную матрицу для нашего примера.

Продолжим вычисление для примера. Подставим известные матрицы в уравнение для вектора коэффициентов $\mathbf{B}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.
Имеем

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1/5 & 0 \\ 0 & 1/10 \end{bmatrix} \cdot \begin{bmatrix} 10 \\ 10 \end{bmatrix}.$$

Перемножим матрицы и получим

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1/5 \cdot 10 & +0 \cdot 10 \\ 0 \cdot 10 & +1/10 \cdot 10 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Две матрицы равны, когда равны их соответствующие элементы, поэтому $b_0=2$, $b_1=1$. Таким образом, мы получили результат, совпадающий с полученным ранее без использования матриц.

Введем еще одно важное понятие: каждой квадратной матрице можно поставить в соответствие некоторое число, называемое *определителем* этой матрицы. Определитель представляет собой алгебраическую сумму всех возможных произведений, в каждое из которых входит по одному элементу от каждой строки и от каждого столбца. Причем знак произведения (+ или —) зависит от положения элементов данного произведения матрицы.

Посчитаем определитель для матрицы системы нормальных уравнений. Здесь возможны два произведения, каждое из которых содержит два сомножителя $5 \cdot 10=50$ и $0 \cdot 0=0$.

Определитель принято обозначать

$$\begin{vmatrix} 5 & 0 \\ 0 & 10 \end{vmatrix} = 50 - 0 = 50$$

или

$$\det \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} = 50$$

(det — определитель от латинского determinant).

Вычислим определитель обратной матрицы

$$\begin{vmatrix} 1/5 & 0 \\ 0 & 1/10 \end{vmatrix} = \frac{1}{50} - 0 = \frac{1}{50}.$$

Можно заметить, что между определителями прямой и обратной матрицы существует в данном случае простое соотношение: они являются взаимнообратными числами. Такое соотношение выполняется и в общем случае.

Определитель может быть любым действительным числом, как положительным, так и отрицательным. Он может оказаться и равным нулю. С последним случаем связаны некоторые особенности свойств матрицы. Матрица, определитель которой равен нулю, не имеет обратной. Такую матрицу называют *особенной*, *вырожденной* или *сингулярной*. Если же определитель матрицы не равен нулю, то матрица называется *неособенной*, *невырожденной* или *несингулярной*.

Например, у следующей матрицы определитель равен нулю и она является вырожденной

$$\det \begin{bmatrix} -1 & +2 \\ +1 & -2 \end{bmatrix} = (-1) \cdot (-2) - (+1) \cdot (+2) = 0.$$

Значит, решение системы нормальных уравнений возможно только тогда, когда матрица невырождена, т. е.

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

Это предполагалось и имело место в нашем примере.

4.5.3.3. Обобщение метода наименьших квадратов на многофакторный линейный случай

Пусть имеется k факторов и известно, что отклик и факторы связаны линейно:

$$y = b_0 x_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k.$$

Выпишем для этого случая матрицы \mathbf{X} , \mathbf{Y} и \mathbf{B}

$$\mathbf{X} = \begin{matrix} \begin{bmatrix} x_{01} & x_{11} \dots & x_{k1} \\ x_{02} & x_{12} \dots & x_{k2} \\ \vdots & \vdots & \vdots \\ x_{0N} & x_{1N} & x_{kN} \end{bmatrix} \\ (N \times k) \end{matrix}; \quad \mathbf{Y} = \begin{matrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \\ (N \times 1) \end{matrix}; \quad \mathbf{B} = \begin{matrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} \\ [(k+1) \times 1] \end{matrix}.$$

Запишем исходную систему линейных уравнений:

$$\mathbf{Y} = \mathbf{XB};$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{01} & x_{11} \dots & x_{k1} \\ x_{02} & x_{12} \dots & x_{k2} \\ \vdots & \vdots & \vdots \\ x_{0N} & x_{1N} \dots & x_{kN} \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}.$$

После преобразований, аналогичных рассмотренным в предыдущем параграфе, приходим к следующей формуле:

$$\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y};$$

$$\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = \left(\begin{bmatrix} x_{01} & x_{02} & \dots & x_{0N} \\ x_{11} & x_{12} & \dots & x_{1N} \\ \vdots & \vdots & & \vdots \\ x_{k1} & x_{k2} & \dots & x_{kN} \end{bmatrix} \cdot \begin{bmatrix} x_{01} & x_{11} & \dots & x_{k1} \\ x_{02} & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & & \vdots \\ x_{0N} & x_{1N} & \dots & x_{kN} \end{bmatrix}^{-1} \right) \times$$

$$\times \begin{bmatrix} x_{01} & x_{02} & \dots & x_{0N} \\ x_{11} & x_{12} & \dots & x_{1N} \\ \vdots & \vdots & & \vdots \\ x_{k1} & x_{k2} & \dots & x_{kN} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

Скалярные произведения удобно представлять в виде сумм, т. е. матрицу системы нормальных уравнений можно записать в следующем виде:

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} \sum x_0^2 & \sum x_0 x_1 & \sum x_0 x_2 & \dots & \sum x_0 x_k \\ \sum x_0 x_1 & \sum x_1^2 & \sum x_1 x_2 & \dots & \sum x_1 x_k \\ \dots & \dots & \dots & & \dots \\ \sum x_0 x_k & \sum x_1 x_k & \sum x_2 x_k & \dots & \sum x_k^2 \end{bmatrix}.$$

Так как суммирование ведется от 1 до N по всему множеству опытов, индекс суммирования мы опустили.

Аналогично $\mathbf{X}^T \mathbf{Y}$ есть вектор сумм произведений:

$$\mathbf{X}^T \mathbf{Y} = \begin{bmatrix} \sum y x_0 \\ \sum y x_1 \\ \sum y x_2 \\ \vdots \\ \sum y x_k \end{bmatrix}.$$

Чтобы получить ответ, т. е. вектор \mathbf{B} , остается обратить матрицу $\mathbf{X}^T \mathbf{X}$ и умножить обратную матрицу на $\mathbf{X}^T \mathbf{Y}$.

Проведем обработку данных табл. 4.21 матричным способом

Таблица 4.21

Матрица планирования 2^{7-4} и коэффициенты

Номер опыта	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	+	+	+	+	+	+	+	+	19,3
2	+	+	-	-	-	-	+	+	23,8
3	+	-	-	+	+	-	-	+	31,3
4	+	-	+	-	-	+	-	+	12,8
5	+	-	-	-	+	+	+	-	32,0
6	+	-	+	+	-	-	+	-	14,0
7	+	+	+	-	+	-	-	-	25,0
8	+	+	-	+	-	+	-	-	30,5
b_j	+23,5875	+1,0625	-5,8125	+0,1875	+3,3125	+0,0625	-1,3125	-1,7875	

$$X = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & -1 & -1 & +1 & +1 & +1 & -1 \\ +1 & -1 & +1 & +1 & -1 & -1 & +1 & -1 \\ +1 & +1 & +1 & -1 & +1 & -1 & -1 & -1 \\ +1 & +1 & -1 & +1 & -1 & +1 & -1 & -1 \end{bmatrix};$$

$$X^T = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & -1 & -1 & +1 & -1 & +1 & +1 & -1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \end{bmatrix};$$

$$X^T X = \begin{bmatrix} 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 \end{bmatrix};$$

$$(X^T X)^{-1} = \begin{bmatrix} 1/8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/8 \end{bmatrix};$$

$$X^T Y = \begin{bmatrix} \sum y_i x_{0i} \\ \sum y_i x_{1i} \\ \sum y_i x_{2i} \\ \sum y_i x_{3i} \\ \sum y_i x_{4i} \\ \sum y_i x_{5i} \\ \sum y_i x_{6i} \\ \sum y_i x_{7i} \end{bmatrix};$$

$$b_0 = \frac{\sum y_i x_{0i}}{8} = \frac{188,7}{8} = 23,5875;$$

$$b_1 = \frac{\sum y_i x_{1i}}{8} = \frac{8,5}{8} = 1,0625;$$

$$b_2 = \frac{\sum y_i x_{2i}}{8} = \frac{-46,5}{8} = -5,8125;$$

$$b_3 = \frac{\sum y_i x_{3i}}{8} = \frac{1,5}{8} = 0,1875;$$

$$b_4 = \frac{\sum y_i x_{4i}}{8} = \frac{26,5}{8} = 3,3125;$$

$$b_5 = \frac{\sum y_i x_{5i}}{8} = \frac{0,5}{8} = 0,0625;$$

$$b_6 = \frac{\sum y_i x_{6i}}{8} = \frac{-10,5}{8} = -1,3125;$$

$$b_7 = \frac{\sum y_i x_{7i}}{8} = \frac{-14,3}{8} = -1,7875.$$

Мы получили те же результаты, что и приведенные в последней строке таблицы 4.21.

Детально рассмотренная процедура МНК в матричной форме показывает, как получаются формулы для коэффициентов регрессии, которые использовались в предыдущих разделах.

Аналогичным путем можно оценить эффекты взаимодействия, входящие в модель. Для этого надо расширить матрицу **X**, включив в нее столбцы взаимодействий. Все остальные операции производятся совершенно аналогично. В векторе **B** появляются при этом элементы, соответствующие эффектам взаимодействий. Расширение матрицы **X** подобным образом называют *линеаризацией*. Это эквивалентно замене эффектов взаимодействия новыми линейными членами. Подобная процедура возможна только тогда, когда все коэффициенты входят в уравнение линейно. Такие уравнения называются *линейными по*

параметрам. В некоторых случаях приходится использовать уравнения, *нелинейные* по параметрам. Примером может служить уравнение Аррениуса

$$y = b_0 e^{b_1 x},$$

которое используется в химической кинетике.

4.5.3.4. Статистический анализ

В предыдущем разделе вначале была продемонстрирована процедура МНК, а затем рассмотрены статистические аспекты. Так же построен и этот раздел.

Перейдем к статистическому анализу в матричной форме. Будем предполагать, что постулаты регрессионного анализа выполняются. Что значит провести статистический анализ? Это значит проверить ряд статистических гипотез: гипотезу об адекватности заданной модели, гипотезы о значимости отдельных коэффициентов регрессии и др. Дальнейшее изложение направлено на прояснение некоторых обстоятельств, определивших вид формул статистического анализа регрессионной модели. Знакомство с элементами алгебры матриц поможет сделать это.

Фундаментальную роль в анализе уравнения регрессии играет матрица,

$$\mathbf{M}^{-1} = (\mathbf{X}^T \mathbf{X} s_{(y)}^2)^{-1},$$

которая называется матрицей *дисперсий-ковариаций*. Прямая матрица \mathbf{M} называется *информационной матрицей Фишера*.

В структуре матрицы дисперсий-ковариаций содержится вся информация о статистических свойствах модели. Провести статистический анализ значит извлечь эту информацию. Для этого прежде всего перейдем от матрицы, обратной к матрице системы нормальных уравнений, к матрице \mathbf{M}^{-1} . Оценка дисперсии воспроизводимости $s_{(y)}^2$ — скаляр; $\mathbf{X}^T \mathbf{X}$ — квадратная матрица. Умножить матрицу на скаляр слева или справа — значит умножить на этот скаляр каждый элемент матрицы.

Полученные таким образом произведения имеют определенный статистический смысл. Так, *на главной диагонали матрицы-произведения стоят оценки дисперсий коэффициентов регрессии, вне главной диагонали расположены оценки ковариаций.*

Чтобы познакомиться с понятием *ковариация*, рассмотрим два произвольных вектор-столбца матрицы \mathbf{X} . Во многих случаях важно знать, сколь сильна линейная связь между этими векторами. *Ковариация является одной из мер такой связи.* Чтобы найти

(так принято сокращенно записывать квадратные матрицы, когда все внедиагональные элементы равны нулю). В этом примере $s^2_{(y)} = 1$ и поэтому дисперсии b -коэффициентов $s^2_{(b_j)} = 0,125$. Таким образом, мы пришли к формуле

$$s^2_{(b_j)} = s^2_{(y)}/N,$$

которая уже фигурировала раньше. Она справедлива для ортогональных планов.

Рассмотрим теперь проверку адекватности линейного уравнения регрессии. Дисперсия адекватности равна

$$s^2_{\text{а.к}} = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N - (k + 1)}.$$

Числитель этого выражения — остаточная сумма квадратов — в матричной форме имеет вид

$$\sum_{i=1}^N (y_i - \hat{y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}^T \mathbf{Y} - \mathbf{B}^T \mathbf{X}^T \mathbf{Y}.$$

Приведем конкретный пример расчета остаточной суммы квадратов разными способами (см. табл. 4.22).

Таблица 4.22

Расчет остаточной суммы квадратов

Номер опыта	x_0	x_1	x_2	y	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
1	+	-	-	95,0	94,5	+0,5	0,25
2	+	+	-	90,0	90,5	-0,5	0,25
3	+	-	+	85,0	85,5	-0,5	0,25
4	+	+	+	82,0	81,5	+0,5	0,25

Линейное уравнение регрессии имеет вид

$$\hat{y} = 88,0 - 2,0x_1 - 4,5x_2.$$

Обычный способ вычисления остаточной суммы дает

$$\sum_{i=1}^4 (y_i - \hat{y}_i)^2 = (0,25 + 0,25 + 0,25 + 0,25) = 1,00.$$

Воспользуемся теперь матричной записью. Для этого придется использовать еще одну матричную операцию — операцию *вычитания*. Разностью двух матриц одинакового размера называется матрица, элементы которой являются разностями соответствующих элементов матриц уменьшаемого и вычитаемого, взятых в том же порядке:

$$Y - \hat{Y} = \begin{bmatrix} 95,0 \\ 90,0 \\ 85,0 \\ 82,0 \end{bmatrix} - \begin{bmatrix} 94,5 \\ 90,5 \\ 85,5 \\ 81,5 \end{bmatrix} = \begin{bmatrix} +0,5 \\ -0,5 \\ -0,5 \\ +0,5 \end{bmatrix}.$$

Транспонирование и перемножение дают

$$(Y - \hat{Y})^T (Y - \hat{Y}) = [+0,5 \ -0,5 \ -0,5 \ +0,5] \begin{bmatrix} +0,5 \\ -0,5 \\ -0,5 \\ +0,5 \end{bmatrix} = \\ = 0,25 + 0,25 + 0,25 + 0,25 = 1,00.$$

Если воспользоваться другим выражением для остаточной суммы квадратов, то

$$Y^T Y - B^T X^T Y = [95,0 \ 90,0 \ 85,0 \ 82,0] \cdot \begin{bmatrix} 95,0 \\ 90,0 \\ 85,0 \\ 82,0 \end{bmatrix} - \\ - [88,0 \ -2,0 \ -4,5] \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 95 \\ 90 \\ 85 \\ 82 \end{bmatrix} = 1,00.$$

$$Y^T Y - B^T X^T Y = [95,0 \ 90,0 \ 85,0 \ 82,0] \cdot \begin{bmatrix} 95,0 \\ 90,0 \\ 85,0 \\ 82,0 \end{bmatrix} - \\ - [88,0 \ -2,0 \ -4,5] \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 95 \\ 90 \\ 85 \\ 82 \end{bmatrix} = 1,00.$$

Здесь мы встретились с необходимостью перемножить три матрицы. Эту операцию можно выполнять либо справа налево, либо слева направо. Сначала перемножается первая пара матриц. А затем матрица-произведение рассматривается как новая матрица и перемножается с третьей. Конечно, размеры всех матриц должны удовлетворять условиям существования произведения, что и выполняется в нашем случае.

Кроме статистических оценок, уже нам известных, введем еще одну оценку — оценку *дисперсии предсказанного значения отклика*. Если имеется адекватное уравнение регрессии, то его можно использовать для предсказания результата какого-нибудь нового опыта в некоторой точке факторного пространства. Для этого достаточно подставить в уравнение координаты этой точки и произвести алгебраические операции. Очевидно точность такого предсказания будет неодинакова в разных точках факторного пространства. Чтобы учесть это различие и вводится дисперсия предсказанного значения- отклика.

Пусть известное уравнение имеет вид $y=b_0+b_1x_1$. Координаты предсказываемой точки задаются вектором $\mathbf{X}_i^T=[1 \ x_i]$. Отсюда следует, что

$$\hat{y}_i = [1 \ x_i] \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}.$$

В рамках предпосылок регрессионного анализа x — неслучайная величина, а b_0 и b_1 — случайные величины, так как они являются функциями результатов эксперимента. Следовательно, \hat{y} — тоже случайная величина, связанная с некоторой суммой двух величин b_0 и b_1 . Дисперсии и ковариация b_0 и b_1 уже известны. Они являются элементами матрицы дисперсий-ковариаций. Для определения дисперсии предсказанного значения отклика $s^2_{\{\hat{y}_i\}}$ можно воспользоваться законом сложения ошибок. Если $y=f(z_1, z_2, \dots, z_k)$, где z_j — независимые случайные величины, входящие в известное уравнение, то

$$s^2_{\{y_i\}} \approx \sum_{j=1}^k \left(\frac{\partial y}{\partial z_j}\right)_i^2 s^2_{\{z_j\}} + \sum_{j=1}^k \sum_{\substack{l=1 \\ j \neq l}}^k \left(\frac{\partial^2 y}{\partial z_j \partial z_l}\right)_i^2 \text{cov}\{z_j, z_l\}.$$

В нашем случае

$$s^2_{\{y_i\}} \approx \left(\frac{\partial y}{\partial b_0}\right)_i^2 s^2_{\{b_0\}} + \left(\frac{\partial y}{\partial b_1}\right)_i^2 s^2_{\{b_1\}} + \left(\frac{\partial^2 y}{\partial b_0 \partial b_1}\right)_i^2 \text{cov}\{b_0, b_1\}$$

или

$$s^2_{\{y_i\}} \approx s^2_{\{b_0\}} + 2x_i \text{cov}\{b_0, b_1\} + x_i^2 s^2_{\{b_1\}}.$$

Последнее соотношение можно записать также в матричной форме:

$$s^2_{\{\hat{y}_i\}} = \mathbf{X}_i^T (\mathbf{X}^T \mathbf{X})^{-1} s^2_{\{y\}} \mathbf{X}_i = [1 \ x_i] \begin{bmatrix} s^2_{\{b_0\}} & \text{cov}\{b_0, b_1\} \\ \text{cov}\{b_0, b_1\} & s^2_{\{b_1\}} \end{bmatrix} \begin{bmatrix} 1 \\ x_i \end{bmatrix}.$$

Произведем вычисления для нескольких опытов из примера предыдущего раздела. Возьмем в качестве первой точки начало

координат — нулевой опыт. В этом случае вектор \mathbf{X}_i имеет вид $\mathbf{X}_i^T = [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$. Матрица

$$\mathbf{M}^{-1} = (\mathbf{X}^T \mathbf{X} s_{(y)}^2)^{-1} = \begin{bmatrix} 1/8 & & & & & & \\ & 1/8 & & & & & \\ & & 1/8 & & & & \\ & & & 1/8 & & & \\ & & & & 0 & & \\ & & & & & 1/8 & \\ & & & & & & 1/8 \\ & & & & & & & 1/8 \end{bmatrix},$$

так как $s_{(y)}^2 = 1$.

Отсюда получим дисперсию предсказанного значения y , который в этой точке равен 23,5875

$$\begin{aligned} s_{(y)}^2 &= [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0] \begin{bmatrix} 1/8 & & & & & & \\ & 1/8 & & & & & \\ & & 1/8 & & & & \\ & & & 1/8 & & & \\ & & & & 0 & & \\ & & & & & 1/8 & \\ & & & & & & 1/8 \\ & & & & & & & 1/8 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \\ &= [1/8 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0] \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = 1/8 = 0,125. \end{aligned}$$

Таким образом, точность предсказания отклика в центре эксперимента оказалась равной дисперсии коэффициентов. Посмотрим, какое значение примет $s_{(y)}^2$ в точке с координатами, равными 2: $\mathbf{X}_9^T = [1 \ 2 \ 2 \ 2 \ 2 \ 2 \ 2]$. Тогда

$$s^2_{(\hat{y}_0)} = [1 \ 2 \ 2 \ 2 \ 2 \ 2 \ 2 \ 2] \begin{bmatrix} 1/8 & & & & & & & \\ & 1/8 & & & & & & \\ & & 1/8 & & & & & \\ & & & 1/8 & & & & \\ & & & & 1/8 & & & \\ & & & & & 1/8 & & \\ & & & & & & 1/8 & \\ & & & & & & & 1/8 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{bmatrix} = 3,625.$$

В этой точке $\hat{y}_0 = 20,0125$, а дисперсия $s^2_{(\hat{y}_0)} = 3,625$. Обратите внимание, как существенно увеличилась дисперсия предсказания при удалении от центра.

Описывая статистический анализ, мы до сих пор не принимали во внимание повторных наблюдений. Перейдем теперь к рассмотрению этого вопроса.

4.5.3.5. Взвешенный метод наименьших квадратов и статистический анализ

Каждый опыт несет некоторую информацию об объекте. Опыты, различающиеся условиями проведения, несут информацию об эффектах факторов, а параллельные опыты позволяют оценить дисперсию воспроизводимости. С ростом числа параллельных опытов растет точность эксперимента и оцениваемые эффекты можно определить с большей надежностью. На практике встречаются различные случаи дублирования опытов. Может оказаться, что к моменту начала эксперимента воспроизводимость опытов известна по предыдущим исследованиям. Так бывает иногда в задачах анализа вещества, когда используется методика с заранее известной ошибкой воспроизводимости. Если предполагать, что в намечаемой серии опытов ошибка не изменится и нет опасности появления грубых наблюдений, то параллельные опыты можно не ставить. Если же мы не располагаем такой информацией по предыдущим исследованиям или считаем наше предположение слишком жестким, тогда приходится дублировать опыты. Сделать это можно по-разному: **в одной точке, в нескольких точках и во всех**. В качестве одной точки выбирается центр плана или некоторая строка матрицы. В других случаях бывает равное число параллельных (равномерное дублирование) или различное (неравномерное дублирование). Последнее часто имеет место потому, что часть опытов может оказаться потерянной: не удался анализ, сломалась установка и т. д. Различные варианты

дублирования опытов приводят к различным вариантам обработки данных.

Начнем с наиболее распространенного случая — *равномерное дублирование*. Если при записи матрицы X не делать различия между параллельными и различными опытами, то число строк в матрице будет равно Nn , где N — число различных опытов; n — число параллельных опытов. Это приведет к некоторым изменениям в системе нормальных уравнений.

Пример 1. Пусть реализован эксперимент 2^2 с двумя параллельными опытами в каждой точке, который дал следующие результаты:

$$X = \begin{bmatrix} +1 & -1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & -1 \\ +1 & +1 & +1 \\ +1 & -1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & -1 \\ +1 & +1 & +1 \end{bmatrix}; \quad Y = \begin{bmatrix} 0,8 \\ 1,3 \\ 1,7 \\ 2,6 \\ 0,6 \\ 1,5 \\ 1,7 \\ 2,7 \end{bmatrix}; \quad B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}.$$

Выпишем матрицу системы нормальных уравнений и найдем оценки коэффициентов

$$X^T X = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ -1 & -1 & +1 & +1 & -1 & -1 & +1 & +1 \\ -1 & +1 & -1 & +1 & -1 & +1 & -1 & +1 \end{bmatrix} \begin{bmatrix} +1 & -1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & -1 \\ +1 & +1 & +1 \\ +1 & -1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & -1 \\ +1 & +1 & +1 \end{bmatrix} = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 8 \end{bmatrix};$$

$$X^T Y = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ -1 & -1 & +1 & +1 & -1 & -1 & +1 & +1 \\ -1 & +1 & -1 & +1 & -1 & +1 & -1 & +1 \end{bmatrix} \begin{bmatrix} 0,8 \\ 1,3 \\ 1,7 \\ 2,6 \\ 0,6 \\ 1,5 \\ 1,7 \\ 2,7 \end{bmatrix} = \begin{bmatrix} 12,9 \\ 4,5 \\ 3,3 \end{bmatrix}.$$

Вычислим значения b -коэффициентов

$$\mathbf{v} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \begin{bmatrix} 1/8 & 0 & 0 \\ 0 & 1/8 & 0 \\ 0 & 0 & 1/8 \end{bmatrix} \begin{bmatrix} 12,9 \\ 4,5 \\ 3,3 \end{bmatrix} = \begin{bmatrix} 1,6125 \\ 0,5625 \\ 0,4125 \end{bmatrix}.$$

Таким образом,

$$\hat{y} = 1,6125 + 0,5625x_1 + 0,4125x_2.$$

В этом варианте расчета различные и параллельные опыты не дифференцировались. Можно поступить иначе, рассматривая матрицу \mathbf{X} как матрицу различных опытов. Тогда для учета информации о параллельных опытах будем использовать так называемую матрицу весов. Она представляет собой квадратную диагональную матрицу \mathbf{P} размера $N \times N$. Элементы главной диагонали равны числу повторных опытов соответствующих строк матрицы \mathbf{X} . Нумерация строк матрицы \mathbf{X} должна совпадать с нумерацией строк матрицы \mathbf{P} .

В нашем примере реализовано четыре различных опыта с двумя параллельными. Поэтому матрица \mathbf{X} и матрица \mathbf{P} имеют вид

$$\mathbf{X} = \begin{bmatrix} +1 & -1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & -1 \\ +1 & +1 & +1 \end{bmatrix}; \quad \mathbf{P} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

Тогда система нормальных уравнений МНК имеет вид

$$(\mathbf{X}^T \mathbf{P} \mathbf{X}) \mathbf{v} = \mathbf{X}^T \mathbf{P} \bar{\mathbf{Y}},$$

где $\bar{\mathbf{Y}}$ — вектор-столбец средних значений по соответствующему числу параллельных опытов. Это усреднение необходимо, чтобы привести в соответствие размеры матриц, входящих в систему нормальных уравнений. Коэффициенты регрессии определяются по формуле

$$\begin{aligned}
 \mathbf{B} &= (\mathbf{X}^T \mathbf{P} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{P} \bar{\mathbf{Y}} = \\
 &= \left(\begin{bmatrix} +1 & +1 & +1 & +1 \\ -1 & -1 & +1 & +1 \\ -1 & +1 & -1 & +1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} +1 & -1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & -1 \\ +1 & +1 & +1 \end{bmatrix} \right)^{-1} \times \\
 &\times \begin{bmatrix} +1 & +1 & +1 & +1 \\ -1 & -1 & +1 & +1 \\ -1 & +1 & -1 & +1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0,70 \\ 1,40 \\ 1,70 \\ 2,65 \end{bmatrix} = \\
 &= \left(\begin{bmatrix} 2 & 2 & 2 & 2 \\ -2 & -2 & 2 & 2 \\ -2 & 2 & -2 & 2 \end{bmatrix} \begin{bmatrix} +1 & -1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & -1 \\ +1 & +1 & +1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 & 2 & 2 & 2 \\ -2 & -2 & 2 & 2 \\ -2 & 2 & -2 & 2 \end{bmatrix} \begin{bmatrix} 0,70 \\ 1,40 \\ 1,70 \\ 2,65 \end{bmatrix} = \\
 &= \begin{bmatrix} 8 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 8 \end{bmatrix}^{-1} \begin{bmatrix} 0,70 \cdot 2 + 1,40 \cdot 2 + 1,70 \cdot 2 + 2,65 \cdot 2 \\ -0,70 \cdot 2 - 1,40 \cdot 2 + 1,70 \cdot 2 + 2,65 \cdot 2 \\ -0,70 \cdot 2 + 1,40 \cdot 2 - 1,70 \cdot 2 + 2,65 \cdot 2 \end{bmatrix} = \\
 &= \begin{bmatrix} \frac{(0,70 + 1,40 + 1,70 + 2,65) \cdot 2}{8} \\ \frac{(-0,70 - 1,40 + 1,70 + 2,65) \cdot 2}{8} \\ \frac{(-0,70 + 1,40 - 1,70 + 2,65) \cdot 2}{8} \end{bmatrix} = \begin{bmatrix} 1,6125 \\ 0,5625 \\ 0,4125 \end{bmatrix}.
 \end{aligned}$$

Как и следовало ожидать, результаты совпадают.

Необходимо обратить внимание на то, что при равномерном дублировании сохраняется ортогональность плана, и матрица нормальных уравнений остается диагональной. При отсутствии параллельных опытов матрица весов становится единичной.

Как же теперь будет выглядеть статистический анализ результатов такого эксперимента? Рассмотрим проверку адекватности модели. При наличии числа повторных опытов n , равного для всех строк плана, дисперсия адекватности равна

$$s_{ад}^2 = \frac{n \sum_{i=1}^N (\bar{y}_i - \hat{y}_i)^2}{N - (k + 1)}.$$

Числитель этого выражения в матричной форме имеет вид

$$n \sum_{i=1}^N (\bar{y}_i - \hat{y}_i)^2 = \bar{Y}^T P \bar{Y} - B^T X^T P \bar{Y}.$$

Повторные опыты накладывают более жесткие условия на проверку адекватности, так как рассчитанный F-критерий увеличивается в n раз и для принятия гипотезы адекватности требуется большее соответствие экспериментальных и расчетных точек. В рассматриваемом примере

$$s_{ад}^2 = 0,0312.$$

Дисперсия воспроизводимости для одинакового числа повторных опытов подсчитывалась, как уже говорилось ранее, по формуле

$$s_{воспр}^2 = \frac{\sum_{i=1}^N \sum_{q=1}^n (y_{iq} - \bar{y}_i)^2}{N(n-1)}.$$

Эта формула справедлива для однородных дисперсий.

Составим таблицу для расчета дисперсии воспроизводимости (табл. 4.23).

Таблица 4.23

Расчет дисперсии воспроизводимости

Номер опыта	Матрица планирования	y'	y''	g	Δy	Δy^2	s_i^2
1	(1)	0,8	0,6	0,70	0,10	0,0100	0,020
2	b	1,3	1,5	1,40	0,10	0,0100	0,020
3	a	1,7	1,7	1,70	0,00	0	0
4	ab	2,6	2,7	2,65	0,05	0,0025	0,005

Проверка показывает, что выборочные дисперсии однородны $G_{эксп} = 0,44$, $G_{табл} = 0,91$, $\alpha = 0,05$, $s_{воспр}^2 = 0,0112$.

Для проверки адекватности линейной модели найдем F-критерий:

$$F = s_{ад}^2 / s_{воспр}^2 = 0,0312 / 0,0112 \approx 2,8.$$

Табличное значение критерия Фишера для числа степеней свободы 1; 4 и 5% уровня значимости (табл. 4.18) равно 7,7. Гипотеза адекватности линейной модели может быть принята.

Осталось проверить значимость b -коэффициентов. Дисперсия оценки b -коэффициентов равна

$$s_{(b_j)}^2 = s_{воспр}^2 / Nn = 0,0014.$$

Дисперсия воспроизводимости $s_{воспр}^2$ деленная на число параллельных опытов n , называется *дисперсией среднего* и обозначается $s_{(g)}^2$.

Отсюда имеем

$$s_{\{b_j\}}^2 = s_{\{\bar{y}\}}^2 / N = 0,0056/4 = 0,0014.$$

Этот же результат получается из матрицы дисперсий-ковариаций $(X^T P X s_{\{\bar{y}\}}^2)^{-1}$ или в нашем примере

$$\begin{bmatrix} 1/8 & 0 & 0 \\ 0 & 1/8 & 0 \\ 0 & 0 & 1/8 \end{bmatrix} 0,0112 = \begin{bmatrix} 0,0014 & 0 & 0 \\ 0 & 0,0014 & 0 \\ 0 & 0 & 0,0014 \end{bmatrix}.$$

Тогда

$$s_{\{b_0\}}^2 = s_{\{b_1\}}^2 = s_{\{b_2\}}^2 = 0,0014.$$

Если при проверке адекватности используется $s_{\{\bar{y}\}}^2$, то числитель F-критерия не нужно умножать на n , поскольку на это число уже поделен знаменатель. Для варианта с равномерным дублированием опытов на практике можно использовать следующую эквивалентную схему обработки результатов, учитывающую усреднение непосредственно.

1. Определим коэффициенты регрессии

$$b_j = \frac{\sum_{i=1}^N \bar{y}_i x_{ji}}{N}.$$

Матрица X в этом случае содержит только отличающиеся вектор-строки, а матрица $P = nE$.

2. Найдем дисперсию адекватности

$$s_{\text{ад}}^2 = \frac{\sum_{i=1}^N (\bar{y}_i - \hat{y}_i)^2}{N - (k + 1)}.$$

3. Оценим дисперсию среднего по строкам

$$s_i^2 = \frac{\sum_{q=1}^n (y_{iq} - \bar{y}_i)^2}{n(n-1)}.$$

4. Проверим гипотезу об однородности дисперсий и после ее принятия найдем общую дисперсию среднего

$$s_{\{\bar{y}\}}^2 = \sum_{i=1}^N s_i^2 / N.$$

5. Затем вычислим дисперсии оценок коэффициентов регрессии

$$s_{\{b_j\}}^2 = s_{\{\bar{y}\}}^2 / N.$$

6. Наконец, проверим гипотезу адекватности модели

$$F = s_{\text{ад}}^2 / s_{\{\bar{y}\}}^2.$$

Применим эту схему обработки результатов к нашему примеру. В табл. 4.24 повторены матрица планирования и средние значения откликов, а также приведены данные, необходимые при проверке адекватности модели. Коэффициенты регрессии $b_0 = 1,6125$; $b_1 = 0,5625$; $b_2 = 0,4125$.

Таблица 4.24

Матрица планирования и результаты опытов

Номер опыта	x_0	x_1	x_2	\bar{y}	\bar{y}	$\bar{y} - \bar{y}$	$(\bar{y} - \bar{y})^2 \cdot 10^4$
1	+1	-1	-1	0,70	0,6375	0,0625	39
2	+1	-1	+1	1,40	1,4625	-0,0625	39
3	+1	+1	-1	1,70	1,7625	-0,0625	39
4	+1	+1	+1	2,65	2,5875	0,0625	39

Дисперсия адекватности $s^2_{ад} = 156 \cdot 10^{-4}$. Дисперсии среднего по строкам (табл. 4.20) $s^2_1 = 0,01$, $s^2_2 = 0,01$, $s^2_3 = 0$, $s^2_4 = 0,0025$. Критерий Кохрена $G = 0,44$, гипотеза об однородности дисперсий принимается.

Общая дисперсия среднего $s^2_{(\bar{y})} = 0,0225/4 = 0,0056$. Дисперсия оценок коэффициентов регрессии $s^2_{(b_j)} = 0,0056/4 = 0,0014$. Проверка гипотезы адекватности модели $F = 0,0156/0,0056 \approx 2,8$.

Пример 2. Пусть реализован план с неравномерным дублированием опытов, в котором первый опыт дублирован дважды. Матрица X имеет вид

$$X = \begin{bmatrix} +1 & -1 & -1 \\ +1 & +1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & +1 \end{bmatrix}.$$

Запишем решение системы нормальных уравнений с учетом весов

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \left(\begin{bmatrix} +1 & +1 & +1 & +1 \\ -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} +1 & -1 & -1 \\ +1 & +1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & +1 \end{bmatrix} \right)^{-1} \times \\ \times \begin{bmatrix} +1 & +1 & +1 & +1 \\ -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0,5(y_{11} + y_{12}) \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}.$$

Выполним указанные действия

$$\mathbf{X}^T \mathbf{P} = \begin{bmatrix} 2 & 1 & 1 & 1 \\ -2 & 1 & -1 & 1 \\ -2 & -1 & 1 & 1 \end{bmatrix}, \quad \mathbf{X}^T \mathbf{P} \mathbf{X} = \begin{bmatrix} 6 & -1 & -1 \\ -1 & 5 & 1 \\ -1 & 1 & 5 \end{bmatrix}.$$

Дублирование одного опыта нарушило ортогональность плана. Применение стандартных формул для подсчета коэффициентов регрессии, используемых в случае ортогональных планов, стало неверным.

Найдем обратную матрицу

$$(\mathbf{X}^T \mathbf{P} \mathbf{X})^{-1} = \frac{1}{28} \begin{bmatrix} 6 & 1 & 1 \\ 1 & 6 & -1 \\ 1 & -1 & 6 \end{bmatrix}.$$

Наконец,

$$\mathbf{X}^T \mathbf{P} \bar{\mathbf{Y}} = \begin{bmatrix} \frac{2(y_{11} + y_{12})}{2} + y_2 + y_3 + y_4 \\ \frac{-2(y_{11} + y_{12})}{2} + y_2 - y_3 + y_4 \\ \frac{-2(y_{11} + y_{12})}{2} - y_2 + y_3 + y_4 \end{bmatrix}.$$

В результате имеем

$$\mathbf{B} = \frac{1}{28} \begin{bmatrix} 6 & 1 & 1 \\ 1 & 6 & -1 \\ 1 & -1 & 6 \end{bmatrix} \cdot \begin{bmatrix} y_{11} + y_{12} + y_2 + y_3 + y_4 \\ -y_{11} - y_{12} + y_2 - y_3 + y_4 \\ -y_{11} - y_{12} - y_2 + y_3 + y_4 \end{bmatrix}.$$

Заметим, что к тому же результату можно придти, используя обычные формулы для нахождения оценок коэффициентов регрессии. Для этого в рассматриваемом примере достаточно сравнить соответствующие матрицы

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 \\ -1 & -1 & +1 & -1 & +1 \\ -1 & -1 & -1 & +1 & +1 \end{bmatrix} \begin{bmatrix} +1 & -1 & -1 \\ +1 & -1 & -1 \\ +1 & +1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & +1 \end{bmatrix} = \begin{bmatrix} 6 & -1 & -1 \\ -1 & 5 & 1 \\ -1 & 1 & 5 \end{bmatrix},$$

$$\mathbf{X}^T \mathbf{Y} = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 \\ -1 & -1 & +1 & -1 & +1 \\ -1 & -1 & -1 & +1 & +1 \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{12} \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} =$$

$$= \begin{bmatrix} y_{11} + y_{12} + y_2 + y_3 + y_4 \\ -y_{11} - y_{12} + y_2 - y_3 + y_4 \\ -y_{11} - y_{12} - y_2 + y_3 + y_4 \end{bmatrix}.$$

Положим, что в ходе экспериментирования производилось дублирование точек в соответствии с матрицей \mathbf{X} и получены результаты, изображаемые вектором \mathbf{Y}

$$\mathbf{X} = \begin{bmatrix} \left. \begin{matrix} +1 & -1 & -1 \\ +1 & -1 & -1 \\ \vdots & \vdots & \vdots \end{matrix} \right\} n_1 \text{ раз} \\ \left. \begin{matrix} +1 & +1 & -1 \\ +1 & +1 & -1 \\ \vdots & \vdots & \vdots \end{matrix} \right\} n_2 \text{ раз} \\ \left. \begin{matrix} +1 & +1 & +1 \\ +1 & +1 & +1 \\ \vdots & \vdots & \vdots \end{matrix} \right\} n_4 \text{ раз} \end{bmatrix}; \quad \mathbf{Y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{41} \\ y_{42} \\ \vdots \\ y_{4n_4} \end{bmatrix},$$

где n_i — число параллельных (дублированных) опытов в i -х условиях $i=1, 2, \dots, N$. Общее число всех опытов будет равно

$$\sum_{i=1}^N n_i = \sum_{i=1}^N (n_i - 1) + N.$$

Перейдем к более лаконичной форме записи условий и результатов эксперимента. Для этого введем матрицу весов \mathbf{P} . Это квадратная диагональная матрица с элементами $p_{ii}=n_i$

$$\mathbf{P} = \begin{bmatrix} n_1 & & 0 \\ & n_2 & \\ 0 & & n_3 \\ & & & n_4 \end{bmatrix}.$$

Эта матрица в совокупности с матрицей \mathbf{X} , содержащей только неповторяющиеся строки, задает условия эксперимента. А его

результаты тогда можно представить в виде вектора \bar{Y} , состоящего из средних наблюдений по дублированным опытам. Для учета различных вариантов дублирования составим таблицу, в которой приведем формулы для различных случаев (табл. 4.25).

Для удобства обозначений условимся в случае равенства всех n_i писать $n_i = \text{const} = n$, а номер единственного опыта, который дублируется, не нарушая общности, будем считать первым. В табл. 4.25 также указаны формулы для случая, когда дисперсия, характеризующая ошибку опыта, находится из независимой серии опытов, результаты которой не используются при вычислении коэффициентов регрессии.

Таблица 4.25

Суммы квадратов и числа степеней свободы при разных вариантах дублирования опытов

Источник рассеяния	Характер дублирования опытов	Суммы квадратов	Число степеней свободы
Ошибка эксперимента	Неравномерный	$\sum_{i=1}^N \sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i)^2$	$\sum_{i=1}^N (n_i - 1)$
	Равномерный	$\sum_{i=1}^N \sum_{q=1}^n (y_{iq} - \bar{y}_i)^2$	$N(n - 1)$
	Дублирование в одной точке $i = 1$	$\sum_{q=1}^{n_1} (y_{1q} - \bar{y}_1)^2$	$n_1 - 1$
	Дублирование в отдельной серии из L опытов	$\sum_{i=1}^L (y_i - \bar{y})^2$	$L - 1$
Неадекватность модели	Неравномерный	$\sum_{i=1}^N n_i (\bar{y}_i - \hat{y}_i)^2$	$N - (k + 1)$
	Равномерный	$n \sum_{i=1}^N (\bar{y}_i - \hat{y}_i)^2$	$N - (k + 1)$
	Дублирование в одной точке $i = 1$	$n_1 (\bar{y}_1 - \hat{y}_1)^2 + \sum_{i=2}^N (y_i - \hat{y}_i)^2$	$N - (k + 1)$
	Дублирование в отдельной серии из L опытов	$\sum_{i=1}^N (y_i - \hat{y}_i)^2$	$N - (k + 1)$

Таким образом, неравномерность дублирования должна учитываться и при регрессионном и при дисперсионном анализе.

Приведенная таблица нуждается в комментариях. Начнем с пояснения содержащихся в ней формул. Для рассмотрения частных случаев, возможных при различных стратегиях дублирования, выразим соответствующие суммы квадратов с помощью обычной формы записи. Прежде всего запишем выражение для остаточной суммы квадратов

$$S_{\text{ост}} = \sum_{i=1}^N \sum_{q=1}^{n_i} (y_{iq} - \hat{y}_i)^2.$$

Разложим эту сумму на составляющие с помощью следующего преобразования:

$$S_{\text{ост}} = \sum_{i=1}^N \sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i + \bar{y}_i - \hat{y}_i)^2,$$

где \bar{y}_i — среднее значение отклика по опытам i -й серии

$$\bar{y}_i = \frac{1}{n_i} \sum_{q=1}^{n_i} y_{iq}.$$

Раскрывая выражение для $S_{\text{ост}}$, будем иметь

$$S_{\text{ост}} = \sum_{i=1}^N \sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i)^2 + \sum_{i=1}^N \sum_{q=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2 + \\ + 2 \sum_{i=1}^N (\bar{y}_i - \hat{y}_i) \sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i).$$

Сумма

$$\sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i),$$

входящая в последнее слагаемое, равна нулю

$$\sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i) = \sum_{q=1}^{n_i} y_{iq} - n_i \bar{y}_i = 0.$$

Поэтому

$$S_{\text{ост}} = \sum_{i=1}^N \sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i)^2 + \sum_{i=1}^N n_i (\bar{y}_i - \hat{y}_i)^2$$

дает разложение $S_{\text{ост}}$ на два слагаемых, первое из которых есть $S_{\text{вопр}}$, а второе - $S_{\text{ад}}$.

С величиной $S_{\text{ост}}$ связано число степеней свободы, равное

$$f_{\text{ост}} = \sum_{i=1}^N n_i - p,$$

где p — число коэффициентов регрессии, входящих в регрессионную зависимость (в линейном случае $p = k + 1$).

Величина

$$S_{\text{воспр}} = \sum_{i=1}^N \sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i)^2$$

связана с числом степеней свободы

$$f_{\text{воспр}} = \sum_{i=1}^N (n_i - 1).$$

Следовательно, дисперсия, характеризующая ошибку эксперимента, есть

$$s_{(y)}^2 = \frac{S_{\text{воспр}}}{f_{\text{воспр}}} = \frac{\sum_{i=1}^N \sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i)^2}{\sum_{i=1}^N (n_i - 1)}.$$

Тот же результат будет, если использовать выражения для частных дисперсий в каждой серии дублированных опытов. Действительно, для i -й серии параллельных опытов можно записать

$$s_i^2 = \frac{S_{\text{воспр } i}}{f_{\text{воспр } i}} = \frac{\sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i)^2}{n_i - 1}.$$

Тогда объединенная оценка дисперсии, характеризующей ошибку опыта, может быть найдена из соотношения

$$s_{(y)}^2 = \frac{f_{\text{воспр } 1}}{\sum_{i=1}^N f_{\text{воспр } i}} s_1^2 + \frac{f_{\text{воспр } 2}}{\sum_{i=1}^N f_{\text{воспр } i}} s_2^2 + \dots + \frac{f_{\text{воспр } N}}{\sum_{i=1}^N f_{\text{воспр } i}} s_N^2.$$

Применение этой формулы оправданно лишь при условии однородности частных дисперсий. Однородность дисперсий может быть проверена в общем случае с помощью критерия Бартлета.

Запишем нашу формулу в виде

$$s_{(y)}^2 = \sum_{i=1}^N s_{\text{воспр } i}^2 f_i / \sum_{i=1}^N f_{\text{воспр } i}.$$

Отсюда следует, что она идентична с предыдущей.

Теперь вернемся к сумме квадратов

$$S_{ад} = \sum_{i=1}^N n_i (\bar{y}_i - \hat{y}_i)^2.$$

Она связана с числом степеней свободы $f_{ад}=N-(k+1)$ или в общем случае $f_{ад}=N-p$. Следовательно, дисперсия, связанная с неадекватностью уравнения регрессии, есть

$$s_{ад}^2 = \frac{S_{ад}}{f_{ад}} = \frac{\sum_{i=1}^N n_i (\bar{y}_i - \hat{y}_i)^2}{N - (k + 1)}.$$

Это позволяет проверить гипотезу о том, что экспериментальные данные не противоречат полученному описанию. Проверка гипотезы производится с помощью дисперсионного отношения Фишера

$$F = s_{ад}^2 / s_{\{y\}}^2.$$

Таким образом, содержащиеся в таблице формулы являются частными случаями одного и того же разложения остаточной суммы квадратов $S_{ост}$. Зная процедуру разложения остаточной суммы квадратов, можно всякий раз получить нужную формулу. Однако на практике более удобно иметь эти формулы в готовом виде.

Наиболее общий случай представляет собой неравномерное дублирование. Чтобы получить здесь значение F-критерия, нужно сумму квадратов, связанную с дисперсией воспроизводимости, разделить на соответствующее ей число степеней свободы и получить знаменатель F-критерия (если построчные дисперсии однородны по критерию Бартлета)

$$s_{\{y\}}^2 = \frac{\sum_{i=1}^N \sum_{q=1}^{n_i} (y_{iq} - \bar{y}_i)^2}{\sum_{i=1}^N (n_i - 1)}.$$

Затем надо взять сумму квадратов, связанную с дисперсией адекватности, поделить на соответствующее ей число степеней свободы и получить числитель

$$s_{ад}^2 = \frac{\sum_{i=1}^N n_i (\bar{y}_i - \hat{y}_i)^2}{N - p}.$$

Аналогично поступают во всех других случаях. Вариант равномерного дублирования уже неоднократно обсуждался. При дублировании в одной строке матрицы (в нашей таблице эта строка имеет индекс 1, хотя ясно, что выбор строк произволен) симметрия нарушается, это находит отражение в формулах. Если экспериментатор

из-за экономии времени ставит параллельные опыты только в одной точке, то часто лучше всего выбирать центр плана, ибо при этом не нарушается ортогональность.

Последний из приведенных в таблице вариантов — дублирование в отдельной серии из L опытов, не входящих в план. Это означает, что параллельные опыты не ставятся, а информация об ошибке опыта черпается из какого-то независимого источника, например, из публикаций предшественников.

Перейдем теперь к рассмотрению требований, которые обычно предъявляются к планам и называются критериями оптимальности.

4.5.3.6. Критерии оптимальности планов

Построение плана эксперимента можно интерпретировать как выбор строк матрицы X , их числа и последовательности проведения.

Этот выбор осуществляется разными способами и соответственно приводит к разным результатам. Это значит, что b -коэффициенты могут быть оценены с разной точностью, что они будут иметь разные ковариации, что предсказанное значение отклика получится с разными дисперсиями и т. д. В зависимости от того, какие требования экспериментатор предъявляет к модели, он может придти к той или иной формулировке требований к матрице X . Формализация этих требований связана с **критериями оптимальности.** ***Критерии оптимальности удобно формулировать в терминах свойств матрицы $M=X^T X$ или матрицы M^{-1} .*** Именно эти матрицы непосредственно связаны с оценками модели и функционально зависят от матрицы X . Так, например, при диагональной матрице M план оказывается ортогональным, т. е. все столбцы матрицы X взаимноортогональны и коэффициенты модели независимы:

$$\text{cov}(b_i, b_j)=0.$$

Таким образом, мы пришли к критерию, который уже был рассмотрен ранее и который является одним из самых существенных для планов, обсуждаемых в нашей книге. Он относится к группе критериев, связанных с оценками свойств коэффициентов. Кроме этой группы критериев будем различать критерии, определяющие предсказательные свойства модели, и критерии, сформулированные без использования матрицы M , такие, как композиционность, возможность разбиения плана на ортогональные блоки, насыщенность и т. д.

Начнем рассмотрение с критериев ***первой группы.*** В их основе лежит концепция совместных эффективных оценок, которая берет

свое начало с работ Р. Фишера, рассматривавшего проблему получения наилучших оценок при обработке экспериментальных данных.

Американский математик Дж. Кифер распространил такое требование на задачи построения планов. Различные критерии этой группы приведены в табл. 4.26. В таблице каждый критерий определяется тремя различными, но эквивалентными способами. Таблица начинается с **критерия D-оптимальности**, с названием которого часто связывается вся концепция оптимального построения планов. Смысл его — **минимизация дисперсии всех коэффициентов регрессии, рассматриваемых как нечто единое, как вектор**. Дисперсию вектора коэффициентов принято называть **обобщенной дисперсией**, которая задается известной нам функцией от матрицы дисперсий-ковариаций — **определителем**. **Чем меньше определитель, тем меньше обобщенная дисперсия.** Как всегда в математике наряду с алгебраическим представлением можно использовать и геометрическое.

Однако в этом случае вместо уже знакомого нам факторного пространства вводится **пространство параметров**, в котором координатные оси задаются значениями коэффициентов регрессии.

Между размерностями факторного пространства и пространства параметров не существует однозначной связи. В случае одного фактора и линейного уравнения регрессии пространство параметров будет двумерным (так как в уравнение входит два коэффициента b_0 и b_1). Но можно представить себе вариант перехода в одномерное пространство, если, например, незначим один из коэффициентов, или в трехмерное пространство, когда от уравнения прямой приходится переходить к параболе. Аналогичное рассуждение имеет место и для большего числа факторов.

Если число факторов и вид уравнения заданы, тогда размерность пространства параметров определяется однозначно. Так, для линейного уравнения с k факторами пространство параметров имеет размерность $k+1$. Каждой точке в таком пространстве соответствует вектор оценок коэффициентов, а определитель матрицы дисперсий-ковариаций пропорционален объему **эллипсоида рассеяния** оценок параметров. Причем центр эллипсоида совмещен с МНК оценкой. Эллипсоид является многомерным аналогом обычного доверительного интервала. Для одномерного пространства параметров он вырождается в отрезок, который и есть доверительный интервал. На плоскости (когда имеется два параметра) получится эллипс. Эллиптическая форма доверительного интервала связана с предпосылкой регрессионного анализа о нормальном распределении.

Геометрическая интерпретация в пространстве параметров характерна для критериев этой группы.

Таблица 4.26

Критерии оптимальности планов для оценок коэффициентов модели

Критерий	Интерпретация		
	алгебраическая	геометрическая	статистическая
D-оптимальность	Минимум определителя матрицы дисперсий-ковариаций $\min \det M^{-1}$	Минимум объема эллипсоида рассеяния оценок параметров	Минимум обобщенной дисперсии
A-оптимальность	Минимум следа матрицы дисперсий-ковариаций $\min \text{tr} M^{-1}$	Минимум суммы квадратов длин осей эллипсоида рассеяния и минимум длины диагональ прямоугольника, описанного около эллипсоида	Минимальная средняя дисперсия оценок коэффициентов
E-оптимальность	Минимум собственного значения матрицы дисперсий-ковариаций $\min \max_k (M^{-1})$	Минимум максимальной оси эллипсоида рассеяния	Отдельные оценки параметров не обладают слишком большими дисперсиями и ковариациями
Минимум дисперсии оценки коэффициентов	Минимум диагонального элемента матрицы дисперсий-ковариаций $\min \max_j s^2(b_j)$	Минимум максимальной проекции оси эллипсоида рассеяния на координатные оси пространства, параметров	Минимум дисперсии оценки коэффициентов
Минимум суммы относительных ошибок оценок	$\min \sum_{j=0}^k \frac{s^2(b_j)}{b_j}$		Минимум суммы относительных ошибок оценок
Ортогональность	Диагональность матрицы дисперсий-ковариаций $\sum_{i=1}^N x_j x_{ij} = \begin{cases} 0 & j \neq i \\ 1 & j = i \end{cases}$	Направление главных осей эллипсоида рассеяния совпадает с направлением координатных осей пространства параметров	Оценки параметров независимы $\text{cov}(b_j, b_u) = 0$ $j \neq u$

Для критерия A -оптимальности матрица X выбирается так, чтобы достигнуть минимума суммы квадратов длин главных осей эллипсоида рассеяния. Алгебраически это соответствует минимуму еще одной

функции матрицы дисперсий-ковариаций, которая называется *следом* и обозначается trM^{-1} (от trace фр.) или spM^{-1} (от Spur нем.). Следом квадратной матрицы называется сумма ее диагональных элементов. Вспомним, что на диагонали матрицы \mathbf{M}^{-1} находятся дисперсии b -коэффициентов. Значит A -оптимальность обеспечивает минимум суммы дисперсий b -коэффициентов без учета их ковариаций и, следовательно, минимум средней дисперсии.

Кроме таких функций от матрицы \mathbf{M}^{-1} как определитель и след для критериев оптимальности можно использовать и другие ее функции. Примером может служить критерий E -оптимальности, минимизирующий максимальное *собственное значение* этой матрицы. Собственное значение принадлежит к характеристикам структуры матрицы. Всегда существует опасность, что эллипсоид рассеяния может получить слишком вытянутую, бананоподобную форму. При этом некоторые b -коэффициенты попадут в неблагоприятные условия. Критерий E -оптимальность позволяет уменьшить эту опасность, поскольку он минимизирует самую длинную ось эллипсоида рассеяния. Подобным образом можно интерпретировать и остальные критерии, приведенные в табл. 4.26.

Перейдем к рассмотрению критериев *второй группы*. Описание этих критериев проводится не на языке пространства параметров, а в более привычных терминах факторного пространства и функции отклика (см. табл. 4.27). G - и Q -критерии связаны с дисперсией предсказания значений отклика.

Таблица 4.27

Критерии оптимальности планов для предсказательных свойств модели

Критерий	Интерпретация	
	алгебраическая	статистическая
G -оптимальность	$\min \max X_i^T \mathbf{M}^{-1} X_i$	Минимум максимального значения дисперсии оценки поверхности отклика
Q -оптимальность	$\min \sum_{i=1}^N X_i^T \mathbf{M}^{-1} X_i$	Минимум средней дисперсии оценки поверхности отклика
Ротатабельность	$X_i^T \mathbf{M}^{-1} X_i = f(\rho)$	Постоянство дисперсии предсказания на равных расстояниях от центра эксперимента
Униформность	$X_i^T \mathbf{M}^{-1} X_i \approx \text{const}$ при $0 \leq \rho \leq 1$	Дисперсия предсказания постоянна в некоторой области вокруг центра эксперимента (например, с единичным радиусом)

G-критерий минимизирует максимальную дисперсию предсказания, Q-критерий — среднюю дисперсию.

Если план *G*-оптимален, то экспериментатор имеет гарантию, что в области планирования не окажется точек, в которых точность оценки поверхности отклика будет слишком низкой.

Критерии ротатабельности и униформности связаны с требованием постоянства дисперсии предсказания на некоторых фиксированных расстояниях от центра эксперимента.

Ротатабельность плана означает, что оценки дисперсии предсказания инвариантны (независимы) относительно вращения координатных осей факторного пространства. Иными словами, дисперсия предсказания не будет зависеть от того, в каком направлении осуществляется движение из начала координат, а зависит только от расстояния между интересующей нас точкой и началом.

Униформность в дополнение к этому требует, чтобы в некоторой окрестности начала координат, обычно внутри сферы единичного радиуса, дисперсия предсказания оставалась приблизительно постоянной.

Кроме этих двух групп критериев, при упоминании которых мы не претендовали на полноту перечисления, существует еще большое число требований, принимаемых во внимание. Укажем некоторые из них.

Наиболее естественное желание экспериментатора — уменьшение числа опытов. ***Минимальное число опытов задается числом коэффициентов модели, а приближение к нему служит мерой насыщенности плана.*** Таким образом, ***насыщенность плана оказывается одним из возможных критериев оптимальности. Именно стремление удовлетворить этому критерию привело к созданию дробных реплик.*** Заметим, что ***дробные реплики одновременно удовлетворяют по крайней мере двум критериям: они ортогональны и насыщены.***

Большое значение имеет требование ***композиционности***, позволяющее разделить эксперимент на части и в случае необходимости последовательно реализовывать одну часть за другой без потери информации. Примером может служить переход от $1/4$ -реплики к полуреплике, а затем к полному факторному эксперименту, если система смешивания оказалась сложной и не позволила выделить интересующие эффекты. Это соответствует последовательному переходу от простой линейной модели к модели с взаимодействиями. То же самое возможно и при переходе от линейных моделей к модели второго порядка и т. д.

Близким к требованию композиционности является требование **разбиения плана на ортогональные блоки**. Композиционность требуется из-за того, что мы не знаем заранее, какой окажется адекватная модель, а разбивать на блоки приходится из-за того, что возможно изменение внешних условий и важно защититься от их влияния.

Список возможных критериев отнюдь не исчерпывается описанным. Но даже если ограничиться перечисленным, то естественно может возникнуть вопрос, зачем нужны простому экспериментатору D -, G - и прочие оптимальности?

Мы уже говорили выше о таких свойствах эксперимента, как ортогональность и ротатабельность.

Ортогональность обеспечивает независимость оценок b -коэффициентов, что очень существенно при интерпретации силы влияния факторов и их взаимодействий, а также, для оценки направления градиента при движении к оптимуму. Поскольку заранее трудно предполагать, в какую сторону будет направлен градиент, то полезно стремиться и к **ротатабельности, обеспечивающей одинаковую точность предсказания в разных направлениях.**

Рассмотрим D -оптимальность. Этот критерий связан с оценкой уравнения регрессии в целом и смысл его становится более понятным, когда речь идет об интерполяционных (описательных) задачах. Здесь уже экспериментатор не стремится изучать влияние каждого фактора в отдельности. **Для него важно получить минимальную обобщенную дисперсию коэффициентов, что обеспечивает достаточные предсказательные свойства модели внутри области эксперимента.** Такие же свойства обеспечиваются и G -критерием. Это происходит не случайно, так как D - и G -критерии во многих случаях эквивалентны. Критерии первой группы отличаются от D -критерия тем, что используют иные свойства вектора оценок коэффициентов регрессии, отличные от обобщенной дисперсии. Могут встретиться постановки задач, требующие того или иного критерия. Однако экспериментатор обычно не может осуществить этот выбор самостоятельно.

Планы, рассматриваемые в настоящей работе, обладают тем свойством, что одновременно являются оптимальными во многих смыслах. Они не только ортогональны и ротатабельны, но еще и D -, G -, A - и E -оптимальны. Такое совпадение имеет место всегда, когда $\mathbf{X}^T\mathbf{X}=\mathbf{N}\mathbf{E}$, где \mathbf{E} — единичная матрица. Матрица \mathbf{X} , состоящая из $+1$ и -1 и удовлетворяющая этому условию, называется матрицей Адамара. Очевидно, матрицы полного факторного эксперимента 2^k и некоторых регулярных дробных реплик принадлежат к этому классу.

Таким образом, рассматриваемые в данной работе планы попутно с теми критериями, на основе которых они строились, удовлетворяют еще и ряду других критериев. Иногда критерии могут оказываться противоречивыми. Тогда выбор наиболее подходящего из них требует большого искусства. Именно искусства, потому что общей теории выбора критериев оптимальности пока не существует

4.5.4. Построение регрессионных моделей

Пусть исследуемый объект определяется выходным параметром y (процент выхода годных изделий, параметр готового изделия или полуфабриката и т. д.), который зависит от совокупности контролируемых входных переменных - факторов

В регрессионном анализе (РА) отклик y представляют в виде суммы результата функционального преобразования $f(z)$ и некоторой случайной ошибки ε (первый постулат РА). Ошибку ε считают нормально распределенной с нулевым математическим ожиданием и постоянной дисперсией σ^2_y , не зависящей ни от z , ни от $f(z)$ (второй постулат РА).

Построение математического описания объекта заключается в нахождении аналитического выражения для функции отклика $f(z)$ и определении дисперсии случайной ошибки σ^2_y . Решение этой задачи осложняется тем, что вид функции чаще всего априори неизвестен.

Полный алгоритм построения регрессионной модели представлен на рис. 4.10. Рассмотрим его основные этапы.

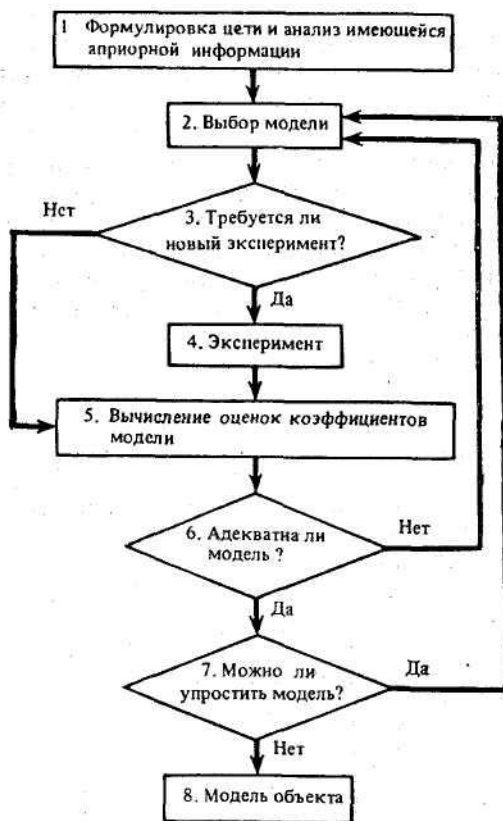


Рис.4.10

4.5.4.1. Выбор модели.

Он заключается в выборе вида аппроксимирующего выражения для неизвестной функции отклика $f(z)$. В РА функцию отклика ищут в виде

$$\hat{y} = b_0 + b_1 x_1 + \dots + b_j x_j + \dots + b_n x_n, \quad (4.88)$$

где b_0, b_1, \dots, b_n — неизвестные коэффициенты; x_1, x_2, \dots, x_n — регрессионные переменные, представляющие собой набор известных линейно независимых функций от физических переменных z_1, \dots, z_l . Например, для двух переменных z_1 и z_2 в качестве регрессионных могут быть взяты

$$x_1 = z_1; x_2 = z_2; x_3 = z_1 z_2 (n = 3);$$

$$x_1 = z_1; x_2 = 1/z_2 (n = 2);$$

$$x_1 = z_1 - z_2; x_2 = \ln z_2; x_3 = z_1^2; x_4 = z_2^2 (n = 4).$$

Такой подход дает целый ряд преимуществ.

Во-первых, большая свобода выбора числа и вида регрессионных переменных позволяет усложнять либо, наоборот, упрощать модель объекта.

Во-вторых, после выбора модели определение функции отклика сводится к нахождению постоянных коэффициентов b_0, b_1, \dots, b_n .

В-третьих, линейность модели (4.88) относительно неизвестных коэффициентов b_0, b_1, \dots, b_n и регрессионных переменных x_1, x_2, \dots, x_n дает возможность широко использовать матричный аппарат линейной алгебры и тем самым добиться максимальной компактности и универсальности при вычислениях.

При выборе числа и вида регрессионных переменных, входящих в (4.88), следует исходить из соображений максимальной простоты модели при одновременном обеспечении приемлемой точности. Большую пользу может принести привлечение имеющейся физической (априорной) информации о природе исследуемого объекта.

Отметим еще один важный момент, который необходимо учитывать при выборе модели. Это область и способ ее использования, т. е. где и как будет использоваться построенная модель. Ответ на эти вопросы часто предопределяет вид модели. Например, если модель предполагают использовать для нахождения экстремума y по методу крутого восхождения, то нет необходимости вводить в модель (4.88) регрессионные переменные, описывающие нелинейные эффекты, т. е. члены вида $z_r z_b, z_k^m$ и т. п.

4.5.4.2. Эксперимент.

Этот этап заключается в наблюдении откликов y_1, y_2, \dots, y_N исследуемого объекта в точках x_1, \dots, x_N (в числе которых могут быть и повторяющиеся) из некоторой области пространства регрессионных переменных. Каждая точка x_i представляет собой конкретный набор x_{i1}, \dots, x_{in} , при котором наблюдается отклик y_i ($i = 1, \dots, N$).

Для того чтобы использовать аппарат матричной алгебры, с каждой точкой пространства регрессионных переменных связывают вектор-строку $x_i = (x_{i0}, x_{i1}, \dots, x_{in})$, в котором для удобства вычислений добавлена еще одна регрессионная переменная x_{i0} , тождественно равная единице. Тогда совокупности данных, полученных в результате эксперимента, могут быть сопоставлены:

матрица наблюдений

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{i1} & x_{i2} & \dots & x_{in} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix}$$

и вектор-столбец откликов

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix}$$

Таким образом, в матрицах X и Y наблюдается построчное соответствие, так что i -я строка отражает результаты проведения i -го опыта.

Отметим, что матрица X может содержать и повторяющиеся строки, что соответствует дублированию опытов при эксперименте.

Существует два подхода к проведению эксперимента: *пассивный* и *активный*.

При *пассивном* эксперименте отсутствует какой-либо определенный порядок в выборе точек проведения эксперимента x . Такой подход оправдан, когда экспериментатор не может поставить опыт в произвольной точке (например, в случае, когда переменные $x_i \dots, x_n$ являются контролируруемыми, но не управляемыми).

При *активном* эксперименте опыты производят в некоторых заранее выбранных точках, образующих план эксперимента. План эксперимента (речь о планировании эксперимента подробно будет идти в разделе б) выбирают в зависимости от выбранной модели и осуществляют по определенной схеме, обладающей некоторыми оптимальными свойствами. С точки зрения введенных матричных обозначений переход к активному эксперименту означает наложение ряда ограничений на матрицу X (матрицу X называют *матрицей планирования*), в результате чего она приобретает определенный вид. Например, в большинстве применяемых планов выполняется условие ортогональности столбцов матрицы планирования:

$$\sum_{i=1}^N x_{ih} x_{il} = 0, \text{ если } h \neq l.$$

В дальнейшем будем отмечать по ходу изложения те особенности, которые вносит в общую схему РА применение математического планирования эксперимента.

4.5.4.3. Вычисление оценок коэффициентов модели.

В РА расчет неизвестных коэффициентов модели (4.88) осуществляют по методу наименьших квадратов (МНК). Его суть заключается в том, что, как мы знаем, коэффициенты определяют из условия минимума суммы квадратов отклонений точек модели от экспериментально наблюдавшихся откликов:

$$Q_{\text{ост}} = \sum_{i=1}^N (\widehat{y}_j - y_i)^2 = \min.$$

Учитывая (4.88), сумма $Q_{\text{ост}}$ может быть записана как

$$Q_{\text{ост}} = \sum_{i=1}^N (b_0 + b_1 x_{i1} + \dots + b_n x_{in} - y_i)^2. \quad (4.89)$$

Таким образом, сумму $Q_{\text{ост}}$ можно рассматривать, как функцию от переменных b_0, b_1, \dots, b_n , поскольку остальные входящие в нее величины есть числа, полученные из эксперимента. Для того чтобы $Q_{\text{ост}}$ была минимальна, необходимо выполнить условие

$$\frac{\partial Q_{\text{ост}}}{\partial b_j} = 0; j = 0, 1, \dots; n. \quad (4.90)$$

Выражение (4.90) представляет собой систему из $(n+1)$ уравнений для определения $(n+1)$ неизвестных коэффициентов b_0, b_1, \dots, b_n . Вследствие линейности модели (4.88) относительно коэффициентов b_j система (4.90) линейна. Ее решение удобно вести в матричном виде. Для этого вводят вектор-столбец коэффициентов:

$$\mathbf{B} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}$$

и вектор-столбец откликов модели

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_N \end{bmatrix}.$$

Тогда

$$\hat{y}_i = x_i \mathbf{B}; \quad \hat{\mathbf{Y}} = \mathbf{X} \mathbf{B}.$$

С использованием введенных обозначений сумма квадратов отклонений $Q_{\text{ост}}$ может быть записана так:

$$Q_{\text{ост}} = (\mathbf{X} \mathbf{B} - \mathbf{Y})^T (\mathbf{X} \mathbf{B} - \mathbf{Y}),$$

где T — знак транспонирования.

Система (4.90) в матричной форме приобретает вид матричного уравнения:

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

откуда искомым вектор коэффициентов

$$\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}, \quad (4.91)$$

где $(\mathbf{X}^T \mathbf{X})^{-1}$ — матрица, обратная матрице $\mathbf{X}^T \mathbf{X}$.

Матрицу $\mathbf{X}^T \mathbf{X}$ называют *информационной*. Легко проверить, что это квадратная матрица порядка $(n + 1)$, симметричная относительно своей главной диагонали. Произведение $\mathbf{X}^T \mathbf{Y}$ представляет собой вектор-столбец, содержащий $(n + 1)$ строк.

Если матрица наблюдений \mathbf{X} содержит большое число повторяющихся строк, т. е. опыты осуществляют с дублированием, то целесообразно изменить схему вычисления вектора коэффициентов \mathbf{B} следующим образом.

Пусть матрица \mathbf{X} содержит N_0 различных строк ($N_0 < N$). Исключим из матрицы \mathbf{X} повторяющиеся строки и оставшуюся матрицу \mathbf{X}_0 будем рассматривать как матрицу различных опытов, каждый из которых производят с числом повторов p_i ($i = 1, \dots, N_0$). Средний отклик в i -м опыте

$$\bar{y}_i = \frac{1}{p_i} \sum_{l=1}^{p_i} y_{il}.$$

Тогда всей совокупности наблюдений можно сопоставить: матрицу различных опытов \mathbf{X}_0 , которая содержит N_0 различных строк, соответствующий ей вектор-столбец \mathbf{Y}_0 средних откликов и матрицу дублирования \mathbf{P} :

$$\mathbf{P} = \begin{bmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \dots & \\ 0 & & & p_{N_0} \end{bmatrix}.$$

Можно показать, что в этом случае

$$\mathbf{B} = (\mathbf{X}_0^T \mathbf{P} \mathbf{X}_0)^{-1} \mathbf{X}_0^T \mathbf{P} \mathbf{Y}_0. \quad (4.92)$$

Результаты вычислений по формулам (4.91) и (4.92) совпадают. Различие состоит в схеме вычислений. При использовании (4.91) не проводят дифференциацию между различными и повторяющимися опытами и поэтому пользуются полной матрицей наблюдений размером $N \times (n + 1)$. Применяя (4.92), за счет дифференциации между различными повторяющимися опытами удается добиться большей компактности при вычислениях, что важно при использовании малых ЭВМ. Особенно нагляден выигрыш в случае равномерного дублирования, когда $p_1 = p_2 = \dots = p_{N_0} = p$. В этом случае матрица дублирования \mathbf{P} в формуле (4.92) сокращается:

$$\mathbf{B} = (\mathbf{X}_0^T \mathbf{X}_0)^{-1} \mathbf{X}_0^T \mathbf{Y}_0 \quad (4.93)$$

т. е. массив числовой информации уменьшается в p раз. Согласно (4.91) и (4.92), не при всякой матрице наблюдений определение коэффициентов регрессионной модели возможно. Действительно, для вычисления вектора \mathbf{B} по (4.91) необходимо, чтобы матрица $\mathbf{X}^T \mathbf{X}$ была обратима, т. е. существовала бы обратная матрица $(\mathbf{X}^T \mathbf{X})^{-1}$. Аналогично, для расчета вектора коэффициентов \mathbf{B} по (4.92) нужно, чтобы была обратима матрица $(\mathbf{X}_0^T \mathbf{P} \mathbf{X}_0)$.

Можно показать, что для существования искомого решения необходимо и достаточно, чтобы была обратима матрица $\mathbf{X}_0^T \mathbf{X}_0$, т. е. успех или неуспех в построении регрессионной модели зависит от того, насколько удачно или неудачно построена матрица различных опытов \mathbf{X}_0 . Известно, что матрица обратима, если ее определитель не равен нулю. Отсюда получаем необходимое и достаточное условие для построения регрессионной модели:

$$|\mathbf{X}_0^T \mathbf{X}_0| \neq 0. \quad (4.93)$$

Если $|\mathbf{X}_0^T \mathbf{X}_0| = 0$, то задача нахождения коэффициентов не имеет решения. Это, в частности, имеет место, когда $N_0 < n + 1$, т. е. число различных опытов меньше числа искомых коэффициентов. Важно отметить, что дублирование опытов в этом случае не имеет значения.

Уточним, какие изменения вносит в описанную схему вычисления коэффициентов модели активный эксперимент (АЭ). С вычислительной точки зрения он сводится к некоторым ограничениям на матрицу наблюдений \mathbf{X} , которую называют *матрицей планирования*. Наиболее существенным из этих ограничений является ортогональность плана. Для ортогонального плана матрица $\mathbf{X}^T \mathbf{X}$ имеет диагональный вид и ее обращение не сложно:

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} A & & & \\ & B & & 0 \\ & & \dots & \\ & 0 & & K \end{bmatrix}; \quad (\mathbf{X}^T \mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{A} & & & \\ & \frac{1}{B} & & 0 \\ & & \dots & \\ & 0 & & \frac{1}{K} \end{bmatrix}.$$

Обычно используют ортонормированные планы, когда

$$\sum_{i=1}^N x_{ik} x_{il} = \begin{cases} 0, & k \neq l, \\ N, & k = l. \end{cases}$$

В этом случае коэффициенты вычисляют по формуле

$$b_j = \frac{1}{N} \sum_{i=1}^N x_{ij} y_i, \quad j = 0, 1, \dots, n.$$

Если АЭ осуществляют с равномерным дублированием, то это не нарушает ортогональности плана и коэффициенты модели рассчитывают по аналогичной формуле

$$b_j = \frac{1}{N_0} \sum_{i=1}^{N_0} x_{ij} \bar{y}_i, \quad j = 0, 1, \dots, n,$$

где \bar{y}_i — средний отклик в i -м опыте.

Таким образом, переход к АЭ существенно упрощает процесс вычисления коэффициентов модели. Кроме того, при АЭ получают диагональную матрицу $\mathbf{X}^T \mathbf{X}$, т. е. наиболее далекую от вырождения (напомним, что вырожденной называют матрицу, определитель которой равен нулю).

Планированию эксперимента должно быть уделено серьезное внимание, так как при неудачном планировании матрица $\mathbf{X}^T \mathbf{X}$ может быть плохо обусловлена (плохо обусловленной называют матрицу, определитель которой очень мал по сравнению с другими значениями). Плохая обусловленность матрицы $\mathbf{X}^T \mathbf{X}$ приводит к тому, что обратная матрица $(\mathbf{X}^T \mathbf{X})^{-1}$ состоит из больших чисел, так как ее члены обратно пропорциональны определителю исходной матрицы. Поскольку дисперсии оценок коэффициентов b_j модели пропорциональны диагональным членам матрицы $(\mathbf{X}^T \mathbf{X})^{-1}$ получают модель, в которой численные значения коэффициентов имеют большую неопределенность. При практическом использовании такая модель имеет слишком низкую точность предсказания.

Рассмотрим модель

$$\widehat{y} = b_0 + b_1z + b_2z^2.$$

Допустим, что для нахождения параметров модели поставили ряд повторных опытов в точках z_1 и z_2 (рис. 4.11).

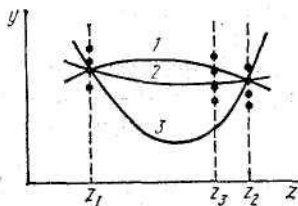


Рис.4.11

При таком планировании не удастся построить требуемую модель, так как располагают данными двух различных опытов, тогда как выбранная модель содержит три неизвестных коэффициента — b_0 , b_1 и b_2 . Тщетность усилий объясняется тем, что пытаются провести параболу через две точки, когда известно, что таких парабол существует бесчисленное множество (например, кривые 1 —3). В данном случае матрица различных опытов X_0 содержит всего две строки, следовательно, при любом числе повторов матрица $X^T X_0$ вырождена.

Очевидно, что необходимо поставить хотя бы один опыт в точке, отличной от z_1 и z_2 . Допустим, что этот опыт (или серия опытов) проведен в точке z_3 . Задача нахождения параметров модели формально стала разрешимой, так как располагают тремя различными опытами. Однако условия второго и третьего опытов различаются незначительно, поэтому следует ожидать большой неопределенности в результатах вычисления коэффициентов модели. Действительно, если кривая 3 не годится в качестве модели, то кривые 1 и 2 по-прежнему хорошо согласуются с экспериментальными данными, тогда как в интервале между точками z_1 и z_3 эти кривые имеют различный вид. Этой геометрически понятной ситуации алгебраически соответствует плохая обусловленность матрицы $X^T X_0$.

Для того чтобы матрица $X^T X_0$ была хорошо обусловленной, нужно выбрать точку z_3 строго посередине между z_1 и z_2 . Целесообразность такого выбора легко понять из геометрических соображений.

В данном примере рассмотрена квадратичная модель. Очевидно, что в случае линейной модели

$$\bar{y} = b_0 + b_1z$$

проведение эксперимента в точках z_1, z_2, z_3 не вызывает трудностей, с которыми пришлось столкнуться в случае квадратичной модели. Следовательно, судить о том, хорош или плох план проведения эксперимента, не имеет смысла, если не имеется в виду конкретная модель, для построения которой используют результаты этого эксперимента.

Проанализируем процесс определения коэффициентов регрессионной модели.

Любое вычислительное устройство с ограниченным объемом памяти работает с конечной погрешностью, обусловленной округлением промежуточных результатов. Эта погрешность может быть особенно большой при обработке результатов пассивного эксперимента, где часто приходится сталкиваться с плохо обусловленной информационной матрицей $\mathbf{X}^T \mathbf{X}$. Если учесть, что для решения задач оптимизации чаще всего используют малые ЭВМ, то уменьшение погрешностей, связанных с округлением, является важной задачей. Для уменьшения погрешностей целесообразно центрировать и нормировать данные эксперимента. Для этого в матрице наблюдений \mathbf{X} и векторе откликов \mathbf{Y} переходят от абсолютных значений величин к их отклонениям от средних значений в эксперименте:

$$x_{ij} \rightarrow x_{ij} - \bar{x}_j; y_i \rightarrow y_i - \bar{y}, \text{ где } \bar{x}_j = \frac{1}{N} \sum_{i=1}^N x_{ij}; \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i.$$

Центрирование данных означает переход к модели

$$\widehat{y} - \bar{y} = b_1 (x_1 - \bar{x}_1) + \dots + b_n (x_n - \bar{x}_n), \quad (4.95)$$

которая не содержит коэффициента b_0 , а следовательно, 1 столбец фиктивных переменных x_{j0} может быть исключен. Центрированные матрицы $\mathbf{X}_c, \mathbf{Y}_c$ используют для вычисления коэффициентов модели (4.95) по формуле (4.92) или (4.93).

Центрирование кроме уменьшения размерности массива данных (информационная матрица $\mathbf{X}_c^T, \mathbf{Y}_c$ имеет порядок n) уменьшает влияние ошибок округления, так как в центрированной модели фигурируют не сами переменные x, y , а их отклонения от средних значений. Например, в экспериментах по диффузионному легированию температуру реактора изменяют в пределах ± 100 °С около средней температуры, которая имеет порядок 1000 °С; центрирование в данном случае облегчает вычисления.

Процесс вычисления облегчается и при переходе к нормированным переменным. Нормирование означает переход к модели

$$\widehat{y} - \bar{y} = b_1 \frac{x_1 - \bar{x}_1}{\Delta_1} + \dots + b_n \frac{x_n - \bar{x}_n}{\Delta_n}, \quad (4.96)$$

где $\Delta_1, \dots, \Delta_n$ — интервалы нормирования.

Нормирование приводит к тому, что все элементы матриц становятся одного порядка, что минимизирует влияние ошибок округления на результат при операциях над матрицами.

4.5.4.4. Проверка модели на адекватность.

Ценность любого математического описания заключается в его предсказующих свойствах. Поэтому за отправную точку при оценке качеств модели выбирают невязку $d = y - \widehat{y}$, т. е. разность между истинным откликом y и тем значением \widehat{y} , которое предсказывает модель.

В качестве меры точности построенной модели в РА используют математическое ожидание квадрата невязки. Если оно во всей рабочей области пространства переменных практически не отличается от дисперсии чистой ошибки σ_y^2 , то построенная модель адекватна исследуемому объекту.

Проверку адекватности модели осуществляют исходя из следующих соображений. Остаточную сумму квадратов $Q_{\text{ост}}$ раскладывают на две составляющие:

$$Q_{\text{ост}} = \sum_{i=1}^{N_0} \sum_{l=1}^{p_i} (y_{il} - \widehat{y}_i)^2 = \sum_{i=1}^{N_0} \sum_{l=1}^{p_i} (y_{il} - \bar{y}_i)^2 + \sum_{i=1}^{N_0} p_i (\bar{y}_i - \widehat{y}_i)^2 = Q_y + Q_{\text{ад}}. \quad (4.97)$$

Здесь индекс i относится к различным опытам ($i = 1, 2, \dots, N_0$), а l — к повторам в i -м опыте ($l = 1, \dots, p_i$).

С каждой из сумм, входящих в (4.97), связывают число степеней свободы f , которое равно числу случайных величин в сумме минус число коэффициентов (констант) в ней, которые вычислены ранее по этим же случайным величинам. Так, $Q_{\text{ост}}$ находят по N значениям откликов y_{il} ; кроме них в сумме фигурируют \widehat{y}_i вычисленные по модели. Но модель содержит $(n + 1)$ коэффициент, которые определены ранее по тем же N значениям откликов y_{il} . Следовательно, число степеней свободы суммы $Q_{\text{ост}}$ равно

$$f_{\text{ост}} = N - (n + 1).$$

Аналогично, суммы $Q_{ост}$ и $Q_{ад}$ имеют число степеней свободы $f_y = N - N_0$ и $f_{ад} = N_0 - (n + 1)$,

так что

$$f_{ост} = f_y + f_{ад} \quad (4.98)$$

Сумма Q_y обусловлена действием чистой ошибки ε . Поэтому выборочная оценка дисперсии $\sigma_{ад}^2$, чистой ошибки

$$s_y^2 = \frac{Q_y}{f_y} = \frac{1}{N - N_0} \sum_{i=1}^{N_0} \sum_{l=1}^{p_i} (y_{il} - \bar{y}_i)^2. \quad (4.99)$$

Сумма $Q_{ад}$ зависит не только от чистой ошибки ε , но и от того, адекватна ли модель описываемому объекту. Поэтому можно рассматривать некоторую дисперсию $\sigma_{ад}^2$, характеризующую качество построенной модели; ее называют *дисперсией адекватности*. Выборочная оценка дисперсии адекватности

$$s_{ад}^2 = \frac{Q_{ад}}{f_{ад}} = \frac{1}{N_0 - (n + 1)} \sum_{i=1}^{N_0} p_i (\bar{y}_i - \hat{y}_i)^2. \quad (4.100)$$

Аналогично вводят остаточную дисперсию $\sigma_{ост}^2$, оценкой которой является

$$s_{ост}^2 = \frac{Q_{ост}}{f_{ост}} = \frac{1}{N - (n + 1)} \sum_{i=1}^{N_0} \sum_{l=1}^{p_i} (y_{il} - \hat{y}_i)^2. \quad (4.101)$$

Если модель адекватна, то дисперсия $\sigma_{ад}^2$ обусловлена чистой ошибкой. Поэтому проверку адекватности модели можно осуществлять как проверку статистической гипотезы

$$H_0: \sigma_{ад}^2 = \sigma_y^2.$$

Как известно, гипотезу о равенстве дисперсий проверяют по F-критерию Фишера. Поэтому для проверки адекватности регрессионной модели следует:

1) вычислить отношение

$$F = s_{ад}^2 / s_y^2,$$

2) по заданному уровню значимости α и числам степеней свободы $f_{ад}$ и f_y в таблице распределения Фишера найти критическое значение критерия F_{α, f_1, f_2} , где $f_1 = f_{ад}$, $f_2 = f_y$;

3) сравнить значения F и F_{α, f_1, f_2} , если $F < F_{\alpha, f_1, f_2}$, то можно признать, что гипотеза об адекватности модели не противоречит опытным данным, если $F \geq F_{\alpha, f_1, f_2}$, то модель неадекватна.

Необходимо отметить, что:

1) построение регрессионной модели невозможно, если число различных опытов меньше числа коэффициентов модели, т. е. $N_0 < n + 1$. В случае, когда $N_0 = n + 1$, модель может быть построена, но проверка ее адекватности неосуществима, так как число степеней свободы суммы $Q_{ад}$ равно нулю. Ценность такой модели при ее дальнейшем использовании сомнительна;

2) если эксперимент проводят без повторов ($N = N_0$), то адекватность модели не может быть проверена на основе данного эксперимента, так как число степеней свободы для оценки σ_y^2 равно нулю. Поэтому построение модели на основе эксперимента без дублирования опытов имеет смысл либо при известной величине σ_y^2 , либо в случае, когда имеется априорная информация об адекватности выбранной модели, т. е. фактически, когда до опыта известен вид функции отклика $f(Z)$ и необходимо определить неизвестные параметры, входящие в эту функцию;

3) построение математического описания исследуемого объекта заключается не только в установлении функции отклика $f(Z)$, но и в нахождении дисперсии случайной ошибки σ_y^2 . Если модель адекватна, то сумма $Q_{ост}$ зависит только от чистой ошибки ϵ , поэтому для вычисления оценки дисперсии σ_y^2 вместо выражения (4.99) следует использовать выражение (4.101), где $Q_{ост}$ можно рассчитать как сумма $Q_y + Q_{3ад}$. Эта оценка является более точной, так как имеет большее число степеней свободы.

4.5.4.5. Упрощение модели.

Пусть построена модель в виде

$$\hat{y} = \bar{y} + b_1(x_1 - \bar{x}_1) + \dots + b_n(x_n - \bar{x}_n).$$

Первый вопрос, который возникает после установления адекватности этой модели,— это вопрос о значимости вклада в модель всех членов выражения, стоящих после \bar{y} , т. е. следует уточнить, насколько оправдано в статистическом смысле усложнение простейшей модели $\hat{y} = \bar{y}$ за счет введения всех последующих членов (проверка значимости регрессии в целом).

Для проверки значимости регрессии общую сумму квадратов отклонений откликов от среднего значения \bar{y} , раскладывают следующим образом:

$$Q_{\text{общ}} = \sum_{i=1}^{N_0} \sum_{l=1}^{p_i} (y_{il} - \bar{y})^2 = \sum_{i=1}^{N_0} \sum_{l=1}^{p_i} (y_{il} - \hat{y}_i)^2 + \sum_{i=1}^{N_0} p_i (\hat{y}_i - \bar{y})^2 = Q_{\text{ост}} + Q_{\text{регр}} \quad (4.102)$$

Соответствующие этим суммам числа степеней свободы связаны соотношением

$$f_{\text{общ}} = f_{\text{ост}} + f_{\text{регр}}; \quad f_{\text{ост}} = N - (n + 1); \quad f_{\text{регр}} = n. \quad (4.103)$$

Сумма $Q_{\text{общ}}$ обусловлена разбросом откликов относительно среднего уровня \bar{y} , ей может быть поставлена в соответствие дисперсия $\sigma_{\text{общ}}^2$, оценка которой $s_{\text{общ}}^2 = Q_{\text{общ}}/f_{\text{общ}}$. Если $\sigma_{\text{общ}}^2$ для адекватной модели незначимо больше $\sigma_{\text{ост}}^2$, то добавление в модель составляющих $b_1(x_1 - \bar{x}_1), \dots, b_n(x_n - \bar{x}_n)$ практически не повышает точность описания исследуемого объекта, следовательно, модель может быть выражена в виде $\hat{y} = \bar{y}$.

Проверку значимости регрессии в целом проводят как проверку статистической гипотезы

$$H_0: \sigma_{\text{общ}}^2 = \sigma_{\text{ост}}^2.$$

Практически эту проверку осуществляют с помощью критерия Фишера в следующем порядке:

1) вычисляют отношение

$$F = s_{\text{общ}}^2/s_{\text{ост}}^2;$$

2) по заданному уровню значимости α и числам степеней свободы $f_{\text{общ}}$ и $f_{\text{ост}}$ в таблице распределения Фишера находят критическое значение F_{α, f_1, f_2} ,

где $f_1 = f_{\text{общ}}, f_2 = f_{\text{ост}}$;

3) сравнивают значения F и F_{α, f_1, f_2} , если $F < F_{\alpha, f_1, f_2}$, то гипотеза H_0 не противоречит опытным данным, т. е. регрессия является незначимой, если $F \geq F_{\alpha, f_1, f_2}$, то регрессия является значимой.

После установления значимости регрессии в целом возникает второй вопрос: нельзя ли упростить модель путем исключения из нее некоторых регрессионных переменных?

Вопрос о возможности исключения из модели переменной x_j может быть решен путем анализа значимости соответствующего коэффициента b_j . Так как величина b_j полученная путем обработки результатов эксперимента, является случайной, зависящей от конкретных значений откликов y_{il} , то значимость b_j можно определить

путем проверки статистической гипотезы $H_0: b_j = 0$. Если значение b_j отличается от нуля незначимо (в статистическом смысле), то переменная x_j может быть исключена из модели.

Для проверки гипотезы $b_j = 0$ необходимо прежде всего оценить меру случайности в определении вектора коэффициентов \mathbf{B} из эксперимента с ограниченным числом наблюдений N . Такой мерой является матрица дисперсий — ковариаций коэффициентов. Диагональные элементы этой матрицы есть дисперсии, а внедиагональные члены — ковариаций коэффициентов b_0, b_1, \dots, b_n . В РА показано, что матрица дисперсий-ковариаций

$$\mathbf{V}(\mathbf{B}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma_y^2. \quad (4.104)$$

При дублировании опытов матрица $\mathbf{V}(\mathbf{B})$ может быть вычислена и так:

$$\mathbf{V}(\mathbf{B}) = (\mathbf{X}_0^T \mathbf{P} \mathbf{X}_0)^{-1} \sigma_y^2, \quad (4.105)$$

где \mathbf{P} — матрица дублирования.

Для равномерного дублирования, когда

$$p_1 = p_2 = \dots = p_n = p,$$

формула (4.105) упрощается:

$$\mathbf{V}(\mathbf{B}) = \frac{1}{p} (\mathbf{X}_0^T \mathbf{X}_0)^{-1} \sigma_y^2. \quad (4.106)$$

Из формул (4.104)—(4.106) следует, что рассеяние оценок истинных коэффициентов модели зависит не только от дисперсии чистой ошибки σ_y^2 , но также и от информационной матрицы $\mathbf{X}^T \mathbf{X}$ (или $\mathbf{X}_0^T \mathbf{X}_0$), т. е. от планирования. В этой связи следует еще раз подчеркнуть достоинства активного эксперимента. Действительно, как уже отмечалось, при АЭ чаще всего используют ортогональные планы, в которых $\mathbf{X}^T \mathbf{X}$ имеет диагональную форму, а следовательно, и $(\mathbf{X}^T \mathbf{X})^{-1}$ также диагональна. Тогда согласно (4.104) в таком эксперименте оценки отдельных коэффициентов модели не зависят друг от друга. Кроме того, при АЭ стремятся к такому выбору матрицы планирования \mathbf{X} , при котором дисперсии коэффициентов модели минимальны (D -оптимальные планы). Таким образом, в АЭ имеется возможность целенаправленно влиять на свойства оценок параметров выбранной модели.

Отметим также, что если σ_y^2 неизвестна, то в формулах (4.104)—(4.106) вместо нее используют S_y^2 , получаемую по (4.101) для адекватной модели.

Проверку значимости каждого коэффициента b_j проводят по критерию Стьюдента в следующем порядке:

1) вычисляют оценку дисперсии коэффициента $s_{b_j}^2$ как соответствующий диагональный элемент матрицы $\mathbf{V}(\mathbf{B})$ по одной из формул (4.104)—(4.106):

2) определяют

$$t = |b_j| \cdot s_{b_j}$$

3) по заданному уровню значимости α и числу степеней свободы $f = N - (n + 1)$ в таблице распределения Стьюдента отыскивают критическое значение $t_{\alpha, f}$

4) если $t < t_{\alpha, f}$, то коэффициент b_j может быть признан незначимым и переменную x_j исключают из модели; если $t \geq t_{\alpha, f}$, то b_j является значимым.

После проверки коэффициентов b_j на значимость и упрощения исходной модели необходимо, в общем случае вернуться к вычислению коэффициентов и проделать все процедуры в соответствии со схемой рис. 4.10 для упрощенной модели.

Отметим, однако, что для этого не нужно заново вычислять матрицу $\mathbf{X}^T \mathbf{X}$. Достаточно из матрицы $\mathbf{X}^T \mathbf{X}$ исходной модели вычеркнуть j -ю строку и j -й столбец, соответствующие исключенной переменной x_j . Вычеркнув строки и столбцы, соответствующие исключенным переменным, получим информационную матрицу $\mathbf{X}^T \mathbf{X}$ для упрощенной модели. Далее необходимо обратить эту матрицу, и так далее до получения новых коэффициентов для упрощенной модели. В том случае, когда матрица $\mathbf{X}^T \mathbf{X}$ исходной модели диагональна (ортогональный план), не нужно пересчитывать коэффициенты упрощенной модели, так как они полностью совпадают с соответствующими коэффициентами исходной модели.

4.5.4.6. Точность регрессионной модели.

После того как получена адекватная регрессионная модель и оценена дисперсия чистой ошибки σ_y^2 , модель может быть использована для предсказания отклика в каждой конкретной точке

$$\mathbf{x}_0 = (1, x_{01}, x_{02}, \dots, x_{0n})$$

пространства переменных.

Следует помнить, что $\hat{y}_0 = \mathbf{x}_0 \mathbf{B}$ — предсказываемое значение функции отклика, т. е. среднего отклика \bar{y}_0 при многократном повторении опытов в точке \mathbf{x}_0 . При этом \hat{y}_0 представляет собой точечную оценку \bar{y}_0 и может не совпадать с истинным значением \bar{y}_0 ,

так как коэффициенты модели b_0, b_1, \dots, b_n определены из эксперимента с ограниченным числом наблюдений и, следовательно, имеют рассеяние, определяемое матрицей дисперсий-ковариаций $\mathbf{V}(\mathbf{B})$.

Соответственно и \hat{y}_0 имеет рассеяние, значение которого определяется дисперсией:

$$D\{\hat{y}_0\} = \mathbf{x}_0 \mathbf{V}(\mathbf{B}) \mathbf{x}_0^T = \mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \sigma_y^2 \mathbf{x}_0^T. \quad (4.107)$$

Используя выражения (4.107), можно построить доверительный интервал для \hat{y}_0 :

$$\hat{y}_0 \pm t_{\alpha; f} \sqrt{D\{\hat{y}_0\}} = \hat{y}_0 \pm t_{\alpha; f} \sqrt{\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} s_y^2 \mathbf{x}_0^T}, \quad (4.108)$$

где s_y^2 — оценка дисперсии ошибки; $t_{\alpha; f}$ — табличное значение критерия Стьюдента при числе степеней свободы, с которым была определена оценка s_y^2 . Истинное значение \bar{y}_0 лежит в пределах этого интервала с вероятностью $(1 - \alpha)$. Если важен не средний отклик \bar{y}_0 , а отклик y_0 при однократном проведении опыта в точке \mathbf{X}_0 , то к дисперсии $D\{\hat{y}_0\}$ следует прибавить дисперсию σ_y^2 , учитывающую независимое действие чистой ошибки ε . Тогда доверительный интервал для y_0 получают из выражения

$$\hat{y}_0 \pm t_{\alpha; f} \sqrt{s_y^2 [\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T + 1]}.$$

В формулах (4.107)—(4.109) \mathbf{x}_0 — вектор-строка соответствующей точки наблюдения. Из формул (4.108), (4.109) следует, что доверительный интервал, т. е. точность предсказания, прежде всего зависит от матрицы $\mathbf{X}^T \mathbf{X}$, т. е. от планирования. Поэтому все замечания, которые были сделаны относительно преимуществ АЭ с точки зрения минимизации рассеяния оценок коэффициентов модели, в равной степени относятся к минимизации доверительного интервала для отклика, предсказываемого моделью.

Кроме того, точность предсказания оказывается различной для разных точек \mathbf{x}_0 . Наилучшая точность достигается в центре эксперимента, на основе которого были вычислены коэффициенты модели (см. центрирование матриц наблюдений \mathbf{X}). По мере удаления от этой точки точность ухудшается.

1. Формулы (4.108), (4.109) справедливы только для области пространства переменных \mathbf{x} , в которой был осуществлен эксперимент по определению коэффициентов регрессионной модели. Механическая экстраполяция на более широкую область пространства переменных \mathbf{x} может привести к большим ошибкам, так как регрессионная модель не

является точным аналитическим выражением функции отклика $f(z)$, а лишь аппроксимирует эту функцию в пределах области проведения эксперимента. При этом нет никакой гарантии в том, что за пределами этой области модель остается адекватной исследуемому объекту. Например, любая непрерывная функция в небольшом интервале изменения аргумента хорошо описывается линейной моделью, однако в более широком интервале линейная аппроксимация уже не является удовлетворительным описанием реальной функции, если эта функция нелинейна;

2. Регрессионная модель теряет реальную ценность для предсказания, если после ее построения происходят внутренние изменения в исследуемом объекте. Такая опасность особенно велика для технологических объектов, в которых, несмотря на все способы стабилизации условий работы обслуживающего персонала и функционирования оборудования, происходят изменения, вызванные различными причинами (смена сырья, изменения, возникшие в продукции на предыдущих технологических операциях, и т. п.).

Рассмотрим пример построения регрессионной модели. Для выяснения влияния времени травления τ мощных сплавных транзисторов перед их герметизацией на коэффициент усиления $h_{21\beta}$ был проверен эксперимент, результат которого представлен в табл. 4.28.

В данном случае модель удобно выбирать визуально, построив зависимость $h_{21\beta} = f(\tau)$ (рис. 4.12). Из рисунка видно, что эта зависимость близка к параболической:

$$h_{21\beta} \approx a + b\tau + c\tau^2.$$

Поэтому выберем модель

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

где

$$x_1 = 0,1\tau; \quad x_2 = (0,1\tau)^2; \quad y = h_{21\beta}.$$

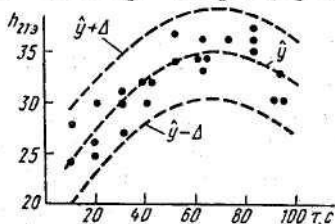


Рис.4.12

Общее число наблюдений в табл. 4.28 $N = 25$, т. е. матрица наблюдений X содержит 25 строк, однако из них различных только $N_0 = 9$.

Таблица 4.28

τ, c	h_{21n}			
10	24	28		
20	25	26	30	
30	27	30	31	
40	30	32	32	
50	34	37		
60	34	33	36	34
70	36			
80	34	36	37	35
90	30	33	30	

Тогда всей совокупности наблюдений можно сопоставить: матрицу различных опытов X_0 , которая содержит $N_0 = 9$ различных строк, соответствующий ей вектор-столбец Y_0 средних откликов и матрицу дублирования P :

$$X_0 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \\ 1 & 8 & 64 \\ 1 & 9 & 81 \end{bmatrix}; \quad Y_0 = \begin{bmatrix} 26 \\ 27 \\ 29,3 \\ 31,3 \\ 35,5 \\ 34,3 \\ 36 \\ 35,5 \\ 31 \end{bmatrix}; \quad P = \begin{bmatrix} 2 & & & & & & & & \\ & 3 & & & & & & & \\ & & 3 & & & & & & \\ & & & 2 & & & & & \\ & & & & 4 & & & & \\ & & & & & 1 & & & \\ & & & & & & 4 & & \\ & & & & & & & 3 & \\ & & & & & & & & 3 \end{bmatrix}$$

Вектор коэффициентов модели согласно (4.92)

$$B = (X_0^T P X_0)^{-1} X_0^T P Y_0.$$

В данном случае

$$X_0^T P X_0 = \begin{bmatrix} 25 & 129 & 831 \\ 129 & 831 & 5991 \\ 831 & 5991 & 45963 \end{bmatrix};$$

$$(X_0^T P X_0)^{-1} = \begin{bmatrix} 0,7 & -0,287 & 0,025 \\ -0,287 & 0,138 & -0,013 \\ 0,025 & -0,013 & 0,001 \end{bmatrix};$$

$$X_0^T P Y_0 = \begin{bmatrix} 794 \\ 4256 \\ 27764 \end{bmatrix}.$$

Окончательно по (4.92) получаем

$$B = \begin{bmatrix} 0,7 & -0,287 & 0,025 \\ -0,287 & 0,138 & -0,013 \\ 0,025 & -0,013 & 0,001 \end{bmatrix} \begin{bmatrix} 794 \\ 4256 \\ 27764 \end{bmatrix} = \begin{bmatrix} 20,2 \\ 4,39 \\ -0,33 \end{bmatrix}.$$

Таким образом, наша модель

$$\hat{y} = 20,2 + 4,39x_1 - 0,33x_2,$$

т.е. коэффициент усиления транзисторов связан с продолжительностью

травления перед герметизацией зависимостью

$$\hat{h}_{213} = 20,2 + 0,439\tau - 0,33 \cdot 10^{-2}\tau^2.$$

Проверим адекватность построенной модели. Для этой цели составим табл. 4.29.

Таблица 4.29

№ опыта i	№ повторения l	Отклик y_{il}	Средний отклик \bar{y}_i	Разность $(y_{il} - \bar{y}_i)^2$	Модель \hat{y}_i	Разность $P_i(\hat{y}_i - \bar{y}_i)^2$	Разность $(y_{il} - \bar{y})^2$
1	1	24	26,0	4	24,2	6,5	60,2
	2	28		4			14,1
2	1	25	27,0	4	27,6	1,1	45,7
	2	26		1			33,2
	3	30		9			3,1
3	1	27	29,3	5,3	30,3	3,0	22,7
	2	30		0,5			3,1
	3	31		2,9			0,6
4	1	30	31,3	1,7	32,4	3,6	3,1
	2	32		0,5			0
	3	32		0,5			0
5	1	34	35,5	2,3	33,8	5,8	5,0
	2	37		2,3			27,4
6	1	34	34,4	0,1	34,5	0,2	5,0
	2	33		1,7			1,5
	3	36		2,9			17,9
	4	34		0,1			5,0
7	1	36	36,0	0	54,6	2,0	17,9
8	1	34	35,5	2,3	34,0	9,0	5,0
	2	36		0,3			17,9
	3	37		2,3			27,4
	4	35		0,3			10,5
9	1	30	31,0	1,0	32,8	9,7	3,1
	2	33		4,0			1,5
	3	30		1,0			3,1
Сумма $\sum y_{il} = 794$				$Q_y = 51,1$		$Q_{ад} = 40,9$	$Q_{общ} = 334$

Заметим, что в шестой столбец заносятся значения откликов, вычисленные по модели

$$\hat{y} = \hat{h}_{213} = 20,2 + 0,439\tau - 0,33 \cdot 10^{-2}\tau^2.$$

а в восьмой — квадраты отклонений экспериментально наблюдавшихся откликов y_{il} от их среднего значения

$$\bar{y} = \frac{1}{N} \sum_{i,l} y_{il} = \frac{1}{25} 794 = 31,76.$$

Для проверки гипотезы об адекватности необходимо вычислить оценки дисперсий σ_y^2 и $\sigma_{ад}^2$. Используем для этого формулы (4.99) и (4.100), взяв из табл. 4.29 $Q_y = 51,1$; $Q_{ад} = 40,9$ и учтя, что числа степеней свободы

$$f_y = N - N_o = 25 - 9 = 16; f_{ад} = N_o - (n + 1) = 9 - 3 = 6.$$

Тогда

$$s_y^2 = Q_y/f_y = 3,2; s_{ад}^2 = Q_{ад}/f_{ад} = 6,8.$$

В соответствии с порядком проверки гипотезы $\sigma_y^2 = \sigma_{ад}^2$ определяем:
а) отношение

$$F = s_{ад}^2/s_y^2 = 2,1,$$

б) критическое значение критерия Фишера из справочных таблиц для

$$f_1 = f_{ап} = 6; f_2 = f_y = 16.$$

Для уровня значимости $\alpha = 0,05$ это значение равно 2,7. Поскольку

$$F < F_{0,05; 6; 16},$$

модель можно считать адекватной

Так как модель оказалась адекватной, скорректируем оценку дисперсии чистой ошибки, применив для этого формулу (4.101)

$$s_y^2 = s_{ост}^2 = Q_{ост}/f_{ост} = (Q_y + Q_{ад}) / (f_y + f_{ад}) = 4,2.$$

Проверим значимость регрессии в целом. Для этого из табл. 4.29 найдем

$$s_{общ}^2 = Q_{общ}/f_{общ} = 13,9.$$

Тогда

$$F = s_{общ}^2/s_{ост}^2 = 3,3.$$

Критическое значение критерия Фишера берут из таблицы Фишера, положив

$$\alpha = 0,05; f_1 = f_{общ} = N - 1 = 24; f_2 = f_{ост} = N - (n + 1) = 22; F_{0,05; 24; 22} = 2,05.$$

Поскольку $F > F_{0,05; 24; 22}$, регрессия в целом значима, т. е. наша модель существенно лучше описывает исследуемый объект, нежели простейшая модель $\hat{y} = \bar{y} = 31,76$.

Попытаемся упростить построенную модель путем исключения отдельных переменных. Согласно (4.105), матрица дисперсий — ковариаций коэффициентов b_j

$$\mathbf{V}(\mathbf{B}) = (\mathbf{X}_0^T \mathbf{P} \mathbf{X}_0)^{-1} s_y^2 = \begin{bmatrix} 0,700 & -0,287 & 0,025 \\ -0,287 & 0,14 & -0,013 \\ 0,025 & -0,013 & 0,001 \end{bmatrix} \cdot 4,2.$$

Отсюда

$$s_{b_0}^2 = 2,9; s_{b_1}^2 = 0,6; s_{b_2}^2 = 0,004.$$

Тогда

$$t_0 = 20,2/\sqrt{2,9} = 11,9; t_1 = 4,39/\sqrt{0,6} = 5,7; t_2 = 0,33/\sqrt{0,004} = 5,2.$$

Зададимся уровнем значимости $\alpha = 0,05$ и для числа степеней свободы $f = N - (n + 1) = 22$, с которым была определена оценка дисперсии σ_y^2 ; из справочных таблиц для t -критерия Стьюдента находим

$$t_{0,05; 22} = 2,07.$$

Поскольку все t_j больше этого критического значения, то все коэффициенты b_j значимы и, следовательно, ни один из членов построенной модели не может быть исключен.

Таким образом, следует принять в качестве рабочей модели

$$\hat{y} = 20,2 + 4,39x_1 - 0,33x_2 = 20,2 + 0,439\tau - 0,33 \cdot 10^{-2} \tau^2.$$

Оценим точность полученной модели, построив для этого доверительные границы для отклика y_0 при проведении однократного опыта в точке x_0 . Эти границы строятся, как $\hat{y}_0 \pm \Delta_0$, где Δ_0 находят в соответствии с формулой (4.109):

$$\Delta_0 = t_{\alpha; f} \sqrt{s_y^2 [\mathbf{X}_0 (\mathbf{X}_0^T \mathbf{P} \mathbf{X}_0)^{-1} \mathbf{X}_0^T + 1]}.$$

Выберем $\alpha=0,05$, т. е. зададимся доверительной вероятностью 0,95. Тогда $t_{0,05;22}=2,07$. Воспользовавшись полученными ранее значениями s_y^2 и $(\mathbf{X}^T \mathbf{P} \mathbf{X}_0)^{-1}$, вычислим доверительные границы для различных точек x_0 , занося результаты в табл. 4.30

Полученные доверительные границы, а также линии регрессии представлены на рис. 4.12. Эти результаты для $\tau = 40$ с, например могут быть интерпретированы так: травление транзистора в течение 40 с перед его герметизацией обеспечивает с доверительной вероятностью 0,95 значение коэффициента усиления $h_{21\beta} = 28 \div 36,8$. Данное утверждение справедливо, если ТП производства транзисторов остается стабильным, т. е. все характеристики процесса, а также сырьё остаются теми же, что и при построения модели.

Таблица 4.30

№ точки	Точка x_0	Отклик модели \hat{y}_0	Δ_0	$\hat{y}_0 + \Delta_0$	$\hat{y}_0 - \Delta_0$
1	$x_{01} = (1, 1, 1)$	24,2	4,8	29,0	19,4
2	$x_{02} = (1, 2, 4)$	27,6	4,5	32,1	23,1
3	$x_{03} = (1, 3, 9)$	30,3	4,4	34,7	25,9
4	$x_{04} = (1, 4, 16)$	32,4	4,4	36,8	28,0
5	$x_{05} = (1, 5, 25)$	33,8	4,4	38,2	29,4
6	$x_{06} = (1, 6, 36)$	34,5	4,4	38,9	30,1
7	$x_{07} = (1, 7, 49)$	34,6	4,4	39,0	30,2
8	$x_{08} = (1, 8, 64)$	34,0	4,4	38,4	29,6
9	$x_{09} = (1, 9, 81)$	32,8	4,7	37,5	28,1

Приложения

Приложение 1

Функция нормального распределения $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z \exp\left(-\frac{z^2}{2}\right) dz$

z	00	01	02	03	04	05	06	07	08	09
0,0	0,000	008	016	024	032	040	048	056	064	072
1	080	088	096	103	111	119	127	135	143	151
2	159	166	174	182	190	197	205	213	221	228
3	236	243	251	259	266	274	281	289	296	303
4	311	318	326	333	340	347	354	362	369	376
0,5	0,383	390	397	404	411	418	425	431	438	445
6	451	458	465	471	478	484	491	497	503	510
7	516	522	528	535	541	547	553	559	565	570
8	576	582	588	593	599	605	610	616	621	627
9	632	637	642	648	653	658	663	668	673	678
1,0	0,683	688	692	697	702	706	711	715	720	724
1	729	733	737	742	746	750	754	758	762	766
2	770	774	778	781	785	789	792	796	799	803
3	806	810	813	816	820	823	826	829	832	835
4	838	841	844	847	850	853	856	858	861	864
1,5	0,866	869	871	874	876	879	881	884	886	888
6	0,890	893	895	867	899	011	031	051	070	090
7	0,9109	127	146	164	181	199	216	233	249	265
8	281	297	312	328	342	357	371	385	399	412
9	426	439	451	464	476	488	500	512	523	534
2,0	0,9545	556	566	576	586	596	606	615	625	634
1	643	651	660	668	676	684	692	700	707	715
2	722	729	736	743	749	756	762	768	774	780
3	786	791	797	802	807	812	817	822	827	832
4	836	840	845	849	853	857	861	865	869	872
2,5	0,9876	879	883	886	889	892	895	898	012	040
6	0,99068	095	121	146	171	195	219	241	264	285
7	307	327	347	367	386	404	422	439	456	473
8	489	505	520	535	549	563	576	590	602	615
9	627	639	650	661	672	682	692	702	712	721

Продолжение

z	00	01	02	03	04	05	06	07	08	09
1,0	0,99730	739	747	755	763	771	779	786	793	800
1	806	813	819	825	831	837	842	848	853	858
2	0,99863	867	872	876	880	885	889	892	896	900
3	0,999033	067	100	132	162	192	221	248	275	301
4	326	350	374	396	418	439	460	480	499	517
3,5	0,999535	552	568	584	600	615	629	643	656	669
6	682	694	705	717	727	738	748	757	767	776
7	784	793	801	809	816	823	830	837	843	849
8	0,999855	861	867	872	877	882	887	891	896	900
9	0,9999038	077	115	151	185	218	251	281	311	339
4,0	0,9999367									

Приложение 2

Функция распределения Стьюдента—Фишера $\Phi_C(\delta) = 2 \int_0^{\delta} f_C(\delta) d\delta$

n	Значение z для P, равного				
	0,5	0,9	0,95	0,99	0,999
2	1,000	6,31	12,7	63,7	637
3	0,816	2,92	4,30	9,92	31,6
4	0,765	2,35	3,18	5,84	13,0
5	0,741	2,13	2,78	4,60	8,61
6	0,727	2,02	2,57	4,03	6,86
7	0,718	1,94	2,49	3,71	5,96
8	0,711	1,90	2,36	3,50	5,40
9	0,706	1,86	2,31	3,36	5,04
10	0,703	1,83	2,26	3,25	4,78
12	0,697	1,80	2,20	3,1	4,49
14	0,694	1,77	2,16	3,01	4,22
16	0,691	1,75	2,13	2,99	4,07
18	0,689	1,74	2,11	2,90	3,96
20	0,688	1,73	2,09	2,86	3,88
25	0,684	1,71	2,06	2,80	3,74
31	0,683	1,70	2,04	2,75	3,65
41	0,681	1,68	2,02	2,70	3,55
61	0,679	1,67	2,00	2,66	3,46
121	0,677	1,65	1,98	2,62	3,37
∞	0,674	1,64	1,96	2,58	3,29

Процентные точки распределения χ^2

$q \backslash f$	99,5 %	97,5 %	95 %	5 %	2,5 %	0,5 %
1	$0,39 \cdot 10^{-4}$	$0,98 \cdot 10^{-3}$	$0,39 \cdot 10^{-2}$	3,841	5,024	7,879
2	0,110	0,050	0,103	5,991	7,378	10,597
3	0,072	0,216	0,352	7,815	9,348	12,838
4	0,207	0,484	0,711	9,488	11,143	14,860
5	0,412	0,831	1,145	11,070	12,832	16,750
6	0,676	1,237	1,635	12,592	14,449	18,548
7	0,989	1,690	2,167	14,067	16,013	20,278
8	1,344	2,180	2,733	15,507	17,535	21,955
9	1,735	2,700	3,325	16,919	19,023	23,589
10	2,156	3,247	3,940	18,307	20,483	25,188
11	2,630	3,816	4,575	19,575	21,920	26,757
12	3,074	4,404	5,226	21,026	23,336	28,300
13	3,565	5,009	5,892	22,362	24,736	29,819
14	4,075	5,629	6,571	23,685	26,119	31,319
15	4,601	6,262	7,261	24,996	27,448	32,801
16	5,142	6,908	7,962	26,296	28,845	34,267
17	5,697	7,564	8,672	27,587	30,191	35,718
18	6,256	8,231	9,390	28,869	31,526	37,156
19	6,844	8,907	10,117	30,144	32,852	38,582
20	7,434	9,591	10,851	31,410	34,170	39,997
21	8,034	10,283	11,591	32,671	35,479	41,401
22	8,643	10,982	12,338	33,924	36,781	42,796
23	9,260	11,688	13,091	35,172	38,076	44,181
24	9,886	12,401	13,848	36,415	39,364	45,558
25	10,520	13,120	14,611	37,652	40,646	46,928
26	11,160	13,844	15,379	38,885	41,923	48,290
27	11,808	14,573	16,151	40,113	43,194	49,645
28	12,461	15,308	16,928	41,337	44,461	50,993
29	13,121	16,047	17,708	42,557	45,722	52,336
30	13,787	16,791	18,493	43,763	46,979	53,672

Критерий исключения промахов

Число наблюдений n	Вероятность, P					
	0,999	0,995	0,99	0,95	0,90	0,80
3	1,414	1,414	1,414	1,414	1,412	1,406
4	1,732	1,730	1,728	1,710	1,689	1,645
5	1,994	1,982	1,972	1,917	1,869	1,791
6	2,212	2,183	2,161	2,067	1,996	1,894
7	2,395	1,344	2,310	2,182	2,093	1,974
8	2,547	2,476	2,431	2,273	2,172	2,041
9	2,677	2,586	2,532	2,349	2,238	2,097
10	2,788	2,680	2,616	2,414	2,294	2,146
11	2,884	2,760	2,689	2,470	2,343	2,190
12	2,969	2,830	2,753	2,519	2,387	2,229
13	3,044	2,892	2,809	2,563	2,426	2,264
14	3,111	2,947	2,859	2,602	2,461	2,297
15	3,171	2,997	2,905	2,638	2,494	2,327
16	3,225	3,042	2,946	2,670	2,523	2,354
17	3,274	3,083	2,983	2,701	2,551	2,380
18	3,320	3,120	3,017	2,728	2,577	2,404
19	3,361	3,155	3,049	2,754	2,601	2,426
20	3,400	3,187	3,079	2,779	2,623	2,447
21	3,436	3,217	3,106	2,801	2,644	2,467
22	3,469	3,245	3,132	2,823	2,664	2,486
23	3,500	3,271	3,156	2,843	2,683	2,504
24	3,529	3,295	3,179	2,862	2,701	2,521

Приложение 5

**Коэффициенты для построения толерантных интервалов
при нормальном распределении**

Число степеней свободы $k=d-1$	$P_{\alpha}=0,9$			$P_{\alpha}=0,95$			$P_{\alpha}=0,99$		
	P			P			P		
	0,9973	0,95	0,9	0,9973	0,95	0,9	0,9973	0,95	0,9
4	6,76	4,18	3,51	8,26	5,11	4,29	12,80	7,92	6,64
5	6,07	3,74	3,14	7,17	4,44	3,72	10,31	6,38	5,35
6	5,60	3,47	2,91	6,50	4,02	3,38	8,91	5,51	4,62
7	5,80	3,27	2,75	6,05	3,74	3,14	8,01	4,95	4,15
8	5,07	3,13	2,63	5,72	3,54	2,97	7,38	4,56	3,83
9	4,89	3,02	2,54	5,48	3,39	2,84	6,91	4,27	3,59
10	4,75	2,94	2,47	5,28	3,26	2,74	6,55	4,05	3,40
12	4,54	2,81	2,36	4,99	3,08	2,59	6,03	3,73	3,13
14	4,39	2,72	2,28	4,78	2,96	2,49	5,67	3,52	2,95
16	4,28	2,65	2,22	4,62	2,86	2,40	5,41	3,35	2,81
18	4,19	2,59	2,17	4,50	2,79	2,34	5,21	3,22	2,70
20	4,11	2,54	2,14	4,39	2,72	2,29	5,05	3,12	2,62
25	3,98	2,46	2,07	4,20	2,61	2,19	4,76	2,94	2,47
30	3,80	2,40	2,02	4,10	2,54	2,13	4,57	2,82	2,37
40	3,78	2,33	1,95	3,94	2,44	2,05	4,31	2,67	2,24
50	3,69	2,28	1,91	3,84	2,37	1,99	4,15	2,57	2,16
60	3,63	2,25	1,89	3,76	2,33	1,96	4,05	2,50	2,10
70	3,59	2,22	1,86	3,70	2,30	1,93	3,96	2,45	2,06
80	3,55	2,20	1,85	3,66	2,27	1,91	3,90	2,41	2,02

Приложение 6

ГРАФИКИ ФУНКЦИЙ ДЛЯ ПОДБОРА ТИПОВ УРАВНЕНИЙ ЭМПИРИЧЕСКИХ ФОРМУЛ

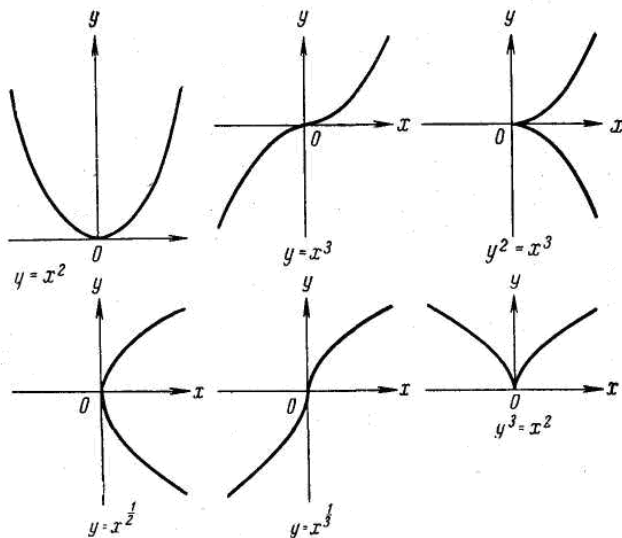


Рис. П.1. Простые степенные зависимости.

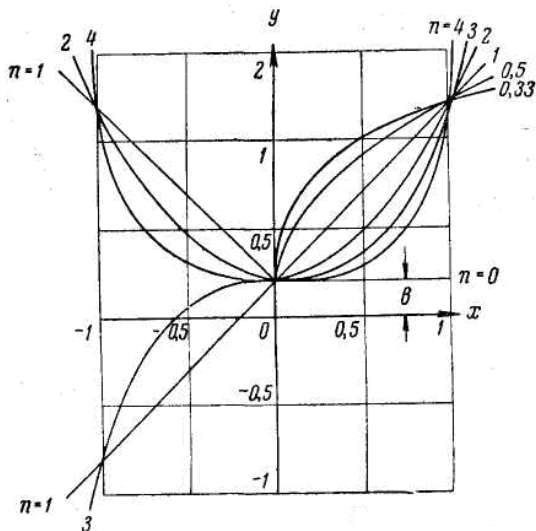


Рис. П.2. Кривые типа $y = ax^n + b$ при $a=1$.

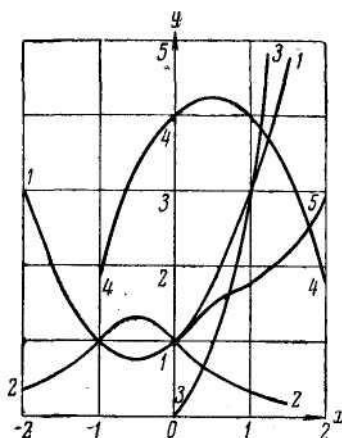


Рис. П.3. График функций:

1—1 — кривая типа $y = a + bx + cx^2$ при $a = b = c = 1$;

2—2 — кривая типа $y = \frac{1}{a + bx + cx^2}$ при $a = b = c = 1$;

3—3 — кривая типа $y = ax + cx^2 + cx^3$ при $a = b = c = 1$; и $x > 0$;

4—4 — кривая типа $y = a + bx + cx$ при $a = 4$; $b = 1$ и $c = -1$;

5—5 — кривая типа $y = ax + bx^2 + cx^3$ при $a < 1,0$, $b < 1,0$ и $c < 1,0$.

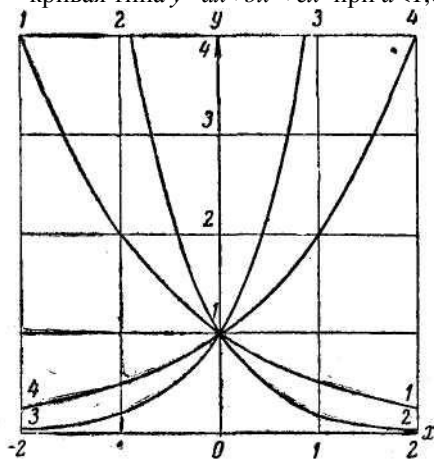


Рис. П.4. Кривые типа $y = a^x$:

1—1 — при $a = 0,5$; 2—2 — при $a = 0,2$; 3—3 — при $a = 5$; 4—4 — при $a = 2$.

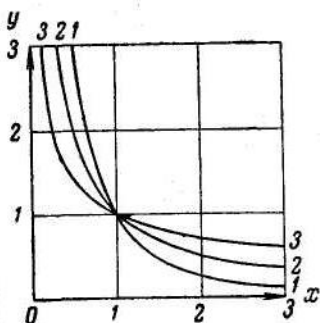


Рис. П.5. Кривые типа $y=x^m$.
 1—1 — при $m=2$; 2—2 — при $m=1$; 3—3 — при $m=0,5$

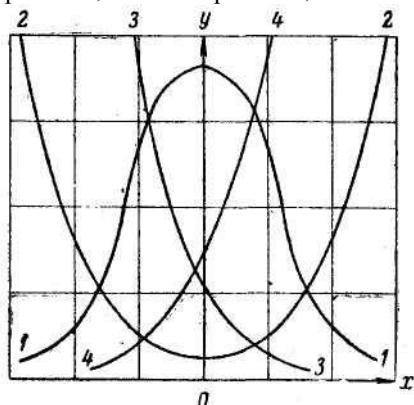


Рис. П.6. График функций:
 1—1 — кривая типа $y = ab^{-xc}$ при $b>0$ и четном значении c ;
 2—2 — кривая типа $y=ab^{xc}$ при $b>0$ и четном значении c ;
 3—3 — кривая типа $y = ab^{-xc}$ при $b>0$ и нечетном значении c ;
 4—4 — кривая типа $y = ab^{xc}$ при $b>0$ и нечетном значении c

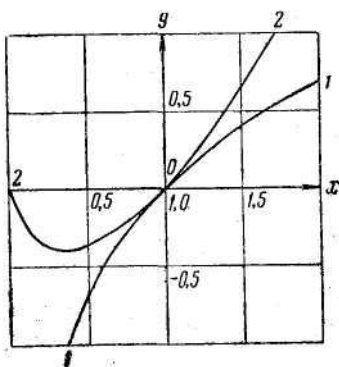


Рис. П.7. График функций:
 1—1 — кривая типа $y = \ln x$; 2—2 — кривая типа $y = x \ln x$.

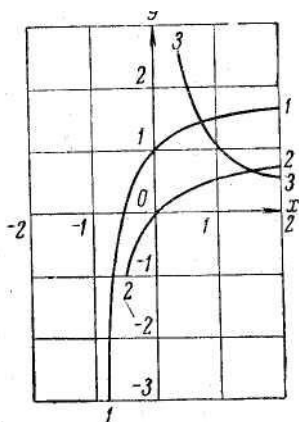


Рис. П.8. График функций:
 1—1 — кривая типа $y = \frac{a+bx}{x} + c$ при $a=b=c=1$;
 2—2 — кривая типа $y = a+bx$ при $a=b=1$;
 3—3 — кривая типа $y = a/x$ при $a=1$.

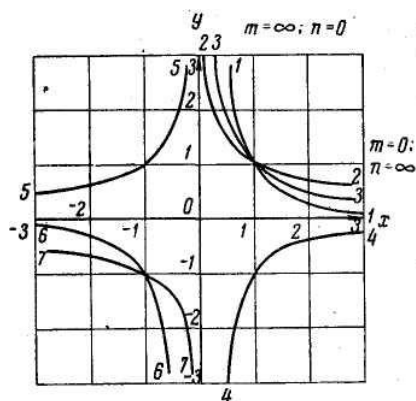


Рис. П.9. Кривые типа $x^m y^n = c$ при $c=1$:
 1—1 — при $m=2, n=1$; 2—2 — при $m=1, n=2$; 3—3 — при $m=n=2$; 4—4 — при $m=3, n=2$; 5—5 — при $m=2, n=3$; 6—6 — при $m>n$, m и n нечетные числа; 7—7 — при $m<n$, m и n — нечетные числа.

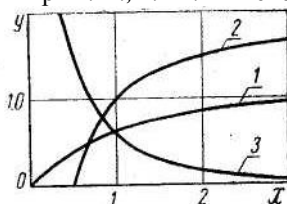


Рис. П.10. График функций:
 1— кривая типа $y=1-e^{-x}$; 2— кривая типа $y=a-(b/x)$ при $a=2$ и $b=1$; 3— кривая типа $y = \frac{1}{e^x - 1}$.

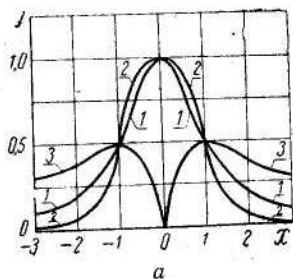
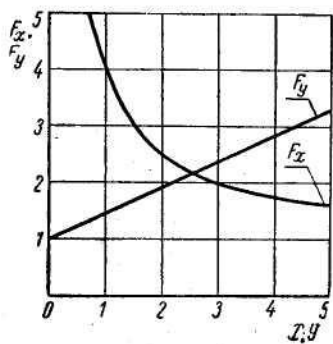
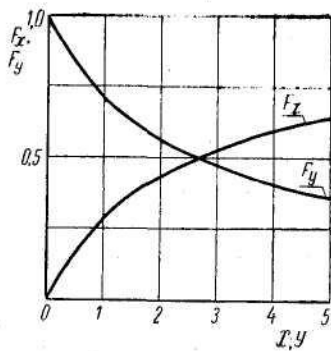


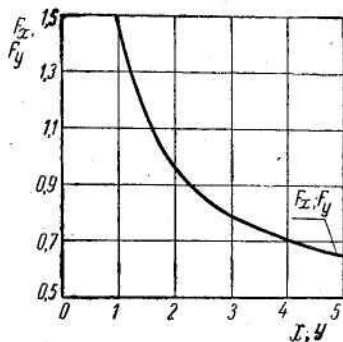
Рис. П.11. График функций $y = \frac{f(x)}{1+x^2}$ при n — положительном:
 1—1 — при $n=2, a=1$; 2—2 — при $n=4, a=1$; 3—3 — при $y = \frac{f(x)}{1+x^2}$.



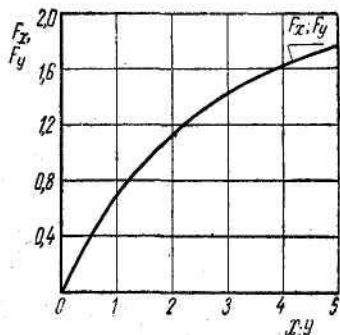
а



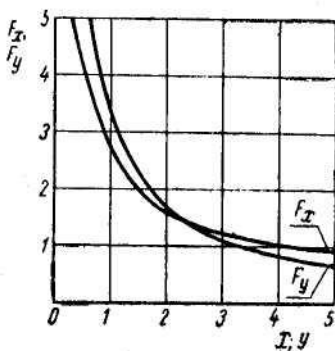
б



в



г



д

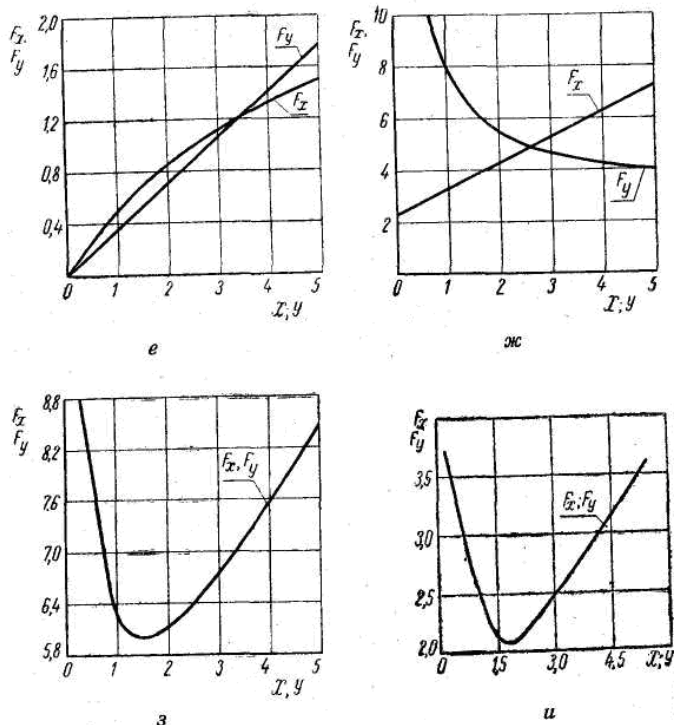


Рис. П.12. Простые типы матричных функционалов:

- | | | |
|----------------|------|--|
| а — функционал | типа | $F_{xy} = \frac{x+y}{x}$; |
| б — функционал | типа | $F_{xy} = \frac{x}{x+y}$; |
| в — функционал | типа | $F_{xy} = \frac{x+y}{xy}$; |
| г — функционал | типа | $F_{xy} = \frac{xy}{x+y}$; |
| д — функционал | типа | $F_{xy} = \frac{x+d}{xy}$; |
| е — функционал | типа | $F_{xy} = \frac{xy}{x+d}$; |
| ж — функционал | типа | $F_{xy} = x + \frac{d}{y}$; |
| з — функционал | типа | $F_{xy} = x+y + \frac{d}{xy}$; |
| и — функционал | типа | $F_{xy} = \frac{x}{y} + \frac{y}{x}$. |

Все графики построены в системе прямоугольных координат с равномерными шкалами. Шкала независимой переменной направлена по оси абсцисс. В подписи к каждому графику указан тип уравнения, соответствующего кривым графика.

При подборе уравнений следует брать участки кривых, наиболее схожие по характеру со сглаженными опытными. Например, если кривая дана в области и положительных, и отрицательных значений независимой переменной, можно брать ее участок в необходимых пределах или в области положительных, или в области отрицательных значений аргумента.

Если условиям схожести удовлетворяют несколько участков кривых различного вида, выбирают формулу наиболее простую или такую, которую можно превратить в полуэмпирическую или полурациональную.

Процентные точки распределения Стьюдента

↙ <i>g</i>					↙ <i>g</i>				
<i>f</i>	10 %	5 %	2 %	1 %	<i>f</i>	10 %	5 %	2 %	1 %
1	6,31	12,71	31,82	63,66	6	1,94	2,45	3,14	3,71
2	2,92	4,30	6,97	9,92	7	1,89	2,36	3,00	3,50
3	2,35	3,18	4,54	5,84	8	1,86	2,31	2,90	3,36
4	2,13	2,78	3,75	4,60	9	1,83	2,26	2,82	3,25
5	2,02	2,57	3,37	4,03	10	1,81	2,23	2,76	3,17
11	1,80	2,20	2,72	3,11	23	1,71	2,07	2,50	2,81
12	1,78	2,18	2,68	3,05	24	1,71	2,06	2,49	2,80
13	1,77	2,16	2,65	3,01	25	1,71	2,06	2,49	2,79
14	1,76	2,15	2,62	2,98	26	1,71	2,06	2,48	2,78

Список литературы

1. *Налимов В. В., Чернова Н. А.* Статистические методы планирования экстремальных экспериментов. М., Наука, 1965.
2. *Зедгендзе И. Г.* Планирование эксперимента для исследования многокомпонентных систем. М., Наука, 1976.
3. Планирование эксперимента в исследовании технологических процессов /Под ред. Э. К. Лецкого. М., Мир, 1977.
4. *Писаренко В. Н., Погорелое А. Г.* Планирование кинетических исследований. М.—Л., Наука, 1969.
5. *Федоров В. В.* Теория оптимального эксперимента. М., Наука, 1971.
6. *Налимов В. В., Голикова Т. И.* Логические основания планирования эксперимента. М., Металлургия, 1976.
7. *Радченко Н.Г.* Устойчивые методы оценивания статистических моделей. К., ПП «Санспарель», 2005.
8. *Бабуров Э.Ф., Куликов Э.Л., Маригодов В.К.* Основы научных исследований. К., Вища школа, 1988.
9. *Веденяпин Г.В.* Общая методика экспериментального исследования и обработки опытных данных. М., Колос, 1973.
10. *Бендат Цок., Пирсол А.* Измерение и анализ случайных процессов. М., Мир, 1974.
11. *Дэниел К.* Применение статистики в промышленном эксперименте. М., Мир, 1979.
12. *Химмельблау Д.* Анализ процессов статистическими методами. М., Мир, 1973.
13. *Статюха Г. А., Ладиев Р. Я.* Байесов подход к анализу сложных ХТС. — Хим. технол., 1973, № 4, с. 43.
14. *Вознесенский В. А., Ковальчук А. Ф.* Принятие решений по статистическим моделям. М., Статистика, 1979.
15. *Адлер Ю. П., Маркова Е. В., Грановский Ю. В.* Планирование эксперимента при поиске оптимальных условий. М., Наука, 1976.
16. *Адлер Ю. П.* Введение в планирование эксперимента. М., Металлургия, 1969.
17. *Рузинов Л. П.* Статистические методы оптимизации химических процессов. М., Химия, 1972.
18. *Егинко В. М.* Организация и проектирование систем автоматизации научно-технических экспериментов. Киев, Наукова думка, 1978.
19. *Островский Г. М., Бережинский Т. А., Беляева А. Р.* Алгоритмы оптимизации химико-технологических процессов. М., Химия, 1978.
20. *Гурский Е. И.* Теория вероятностей с элементами математической статистики. М., Высшая школа, 1971.

Научно-учебное издание

Кононюк Анатолий Ефимович

**ОСНОВЫ НАУЧНЫХ
ИССЛЕДОВАНИЙ**
(Общая теория эксперимента)

Книга 1

Руководитель издательских проектов Кривенко А.А.
Оригинал-макет изготовлено издательством «КНТ»
Редакция авторская
Ответственный за выпуск

Подписано в печать 00.00.2011 г.
Формат 60x84/16.
Усл. печ. л. 18,5.
Тираж экз. Заказ №

Издательство «КНТ»
04210, г.Киев, пр. Героев Сталинграда, 8, корпус 8, оф. 1.
Тел./факс (044) 581-21-38.
E-mail: knt2012@ukr.net

Свидетельство субъекта издательского дела: ДК № 581 от 03.08.2001