МАТЕМАТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Методические указания к курсовой работе для студентов бакалавриата направления 15.03.04

САНКТ-ПЕТЕРБУРГ
2019

Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
Санкт-Петербургский горный университет

Кафедра автоматизации технологических процессов и производств

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Методические указания к курсовой работе для студентов бакалавриата направления 15.03.04

САНКТ-ПЕТЕРБУРГ
2019

УДК 519.242 (073)

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ: Методические указания к курсовой работе / Санкт-Петербургский горный университет. Сост. *Н.В. Васильева.* СПб, 2019. 45 с.

Содержат информацию о порядке проведения работ, об основных требованиях к содержанию разделов, объему, а также процедуре защиты и критериях оценки курсовой работы.

Методические указания предназначены для студентов бакалавриата направления 15.03.04 «Автоматизация технологических процессов и производств».

Научный редактор профессор *В.Ю. Бажин*

Рецензент канд. техн. наук *В.В. Васильев* (ООО «ТОМС инжиниринг»)

ВВЕДЕНИЕ

При исследовании технических систем могут использоваться теоретические и эмпирические методы познания. Каждое из этих направлений обладает относительной самостоятельностью, имеет свои достоинства и недостатки. В общем случае, теоретические методы в виде математических моделей позволяют описывать и объяснять взаимосвязи элементов изучаемой системы или объекта в относительно широких диапазонах изменения переменных величин. Однако при построении теоретических моделей неизбежно введение каких-либо ограничений, допущений, гипотез и т.п. Поэтому возникает задача оценки адекватности (достоверности) полученной модели реальному процессу или объекту. Для этого проводится экспериментальная проверка разработанных теоретических моделей.

Практика является решающей основой научного познания. В ряде случаев именно результаты экспериментальных исследований дают толчок к теоретическому обобщению изучаемого явления. Экспериментальное исследование дает более точное соответствие между изучаемыми параметрами. Однако не следует преувеличивать результаты экспериментальных исследований, которые справедливы только в пределах условий проведенного эксперимента.

Таким образом, теоретические и экспериментальные исследования дополняют друг друга и являются составными элементами процесса познания окружающего нас мира.

Как правило, результаты экспериментальных исследований нуждаются в определенной математической обработке. В настоящее время процедура обработки экспериментальных данных достаточно хорошо формализована, и исследователю необходимо только ее правильно использовать. Круг вопросов, решаемых при обработке результатов эксперимента, не так уж велик. Это вопросы подбора эмпирических формул и оценка их параметров, вопросы оценки истинных значений измеряемых величин и точности измерений, вопросы исследования корреляционных зависимостей и некоторые другие.

1. ПОРЯДОК ПОДГОТОВКИ, ОФОРМЛЕНИЯ И ЗАЩИТЫ КУРСОВОЙ РАБОТЫ

Курсовая работа по дисциплине «Математические методы обработки данных» выполняется студентами направления подготовки 15.03.04 – «Автоматизация технологических процессов и производств» (уровень бакалавриата) в четвертом семестре. В ходе выполнения курсовой работы студенты: самостоятельно прорабатывают теоретическую часть вопроса; производят поиск решений, поставленных в работе задач; посещают консультации, проводимые преподавателем в целях получения разъяснений; формируют пояснительную записку, а также готовят презентацию работы к защите курсовой работы перед комиссией.

При этом необходимо учитывать, что для получения допуска к защите перед комиссией, студенту необходимо сдать готовую курсовую работу преподавателю не позднее, чем за неделю до предполагаемой даты защиты.

Основным документом курсовой работы по дисциплине «Математические методы обработки данных» является пояснительная записка. В ней кратко и четко необходимо раскрыть творческий замысел работы, обоснование методов расчета, сами расчеты, анализ результатов, выводы.

Пояснительная записка должна содержать: титульный лист, лист задания для курсовой работы, содержание, введение, основную часть, выводы, приложения, список литературы.

Объем пояснительной записки должен быть 20-25 листов.

1.1. ПОЛУЧЕНИЕ ЗАДАНИЯ ДЛЯ КУРСОВОЙ РАБОТЫ

Первым шагом при выполнении курсовой работы является получение задания. Каждый студент получает индивидуальное задание. При выдаче задания преподаватель заполняет лист задания, и студент расписывается в ведомости о его получении. Студентам необходимо сохранить выданный преподавателем лист задания, так как он подшивается к пояснительной записке (вторая страница пояснительной записки, сразу за титульным листом). Внешний вид листа задания показан на рисунке 1.1.

Министерство науки и высшего образования Российской федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

Санкт-Петербургский горный университет

УТВЕРЖДАЮ
Заведующий кафедрой
______ /___________/
(подпись) (Ф.И.О.)
"___"__________20 _ г.

Кафедра автоматизации технологических процессов и производств

КУРСОВАЯ РАБОТА

по дисциплине
«   Математические методы обработки данных   »
(наименование учебной дисциплины согласно учебному плану)

ЗАДАНИЕ

студенту группы ____________   ___________________
(шифр группы)   (Ф.И.О.)

1. Тема работы ____
____
2. Исходные данные к работе ____________________________________
____
3. Содержание пояснительной записки ____________________________
____
4. Перечень графического материала ____________________________
____
5. Срок сдачи законченной работы ____ _____________ 20 _ г.

Руководитель работы __________ __________ /_______________/
(должность) (подпись) (Ф.И.О.)

Дата выдачи задания: ____ _____________ 20 __ г.

Рис. 1. Пример листа с заданием для курсовой работы

1.2. ОБЩИЕ ТРЕБОВАНИЯ К ОФОРМЛЕНИЮ ПОЯСНИТЕЛЬНОЙ ЗАПИСКИ

1. Пояснительная записка должна быть написана на одной стороне листа бумаги формата А4 с размерами 297×210 мм. По всем сторонам листа должны быть оставлены поля. Размер левого поля 30 мм, правого – 15 мм, верхнего и нижнего – 20 мм. Шрифт – Times New Roman, кегль 12. Междустрочный интервал – 1,5.

2. Листы пояснительной записки нумеруют арабскими цифрами, проставляя их по центру нижнего поля листа без знаков препинания. На первом листе (титульный лист) номер не ставят. Нумерация листов должна быть сквозной. Рисунки и таблицы, расположенные на отдельных листах, приложения и список литературы включают в общую нумерацию листов.

3. Содержание пояснительной записки делится на разделы, подразделы. Каждый раздел (подраздел) должен иметь наименование, соответствующее его содержанию.

Наименование раздела должно быть кратким и записано в виде заголовка прописными буквами. Точку в конце наименования не ставят. Разделы должны быть пронумерованы арабскими цифрами (нумеруются только разделы основной части пояснительной записки). После номера ставят точку.

Подразделы в пределах раздела должны быть пронумерованы тоже арабскими цифрами. Номер состоит из номеров раздела и подраздела, разделенных точкой. После номера ставят точку.

4. Титульный лист пояснительной записки должен быть выполнен по форме, утвержденной методическим советом Университета.

5. Содержание помещают на втором листе пояснительной записки. В нем последовательно перечисляют все разделы пояснительной записки с указанием номеров листов, на которых они помещены.

6. Во введении необходимо обосновать значимость математической обработки информации для решения задач промышленности (расчет технологических процессов и оборудования применительно к действующим и проектируемым производствам; исследование процесса с целью оптимального управления).

7. В основной части пояснительной записки должна содержаться методика обработки данных, а также расчетная часть.

8. В выводах необходимо указать, какие инженерные задачи могут быть решены с помощью полученной математической модели. В случае неадекватности получаемой модели дать рекомендации для получения адекватной модели.

9. Список литературы (указываются только источники, на которые имеются ссылки в тексте) должен включать фамилию и инициалы автора, заглавие книги, место издания, издательство, год выпуска, количество страниц.

10. В тексте дают ссылки на использованные источники литературы: в квадратных скобках указывают порядковый номер источника по мере появления его в тексте и в списке литературы.

11. Иллюстрированный материал, таблицы и прочее могут быть оформлены как по тексту, так и в виде приложений, на которые должны быть даны ссылки в тексте пояснительной записки.

Каждое приложение должно начинаться с нового листа, иметь тематический заголовок и в правом верхнем углу над заголовком слово «Приложение», без кавычек.

Если в записке больше одного приложения, их нумеруют арабскими цифрами без знака №, например: ПРИЛОЖЕНИЕ 1. Все приложения должны иметь сквозную нумерацию в пределах всей записки.

Курсовая работа должна быть надлежащим образом скреплена. Это предполагает, что при многократном обращении к ней, она должна сохранять свою целостность, то есть не рассыпаться на отдельные листы. Поэтому работы, листы которых скреплены обычной скрепкой, либо вообще не скреплены, а просто вложены в полиэтиленовый файл или папку, приниматься не будут.

1.3. ПОРЯДОК ЗАЩИТЫ КУРСОВОЙ РАБОТЫ

Общее руководство и контроль за ходом выполнения курсовой работы осуществляет научный руководитель – преподаватель соответствующей дисциплины.

На время выполнения курсовой работы составляется расписание консультаций. Необходимо правильно распределить и продуктивно использовать время, отведенное для непосредственного выполнения курсовой работы и подготовки ее к защите. Студент обязан регулярно (согласно графику консультаций) информировать научного руководителя о ходе выполнения курсовой работы, представляя ему для ознакомления и проверки текста отдельные разделы и подразделы в рамках утвержденной программы выполнения работы. В случае серьезных замечаний научного руководителя текст основательно перерабатывается. В исправленном виде в полном объеме и вместе с иллюстрациями он снова подается на проверку руководителю в установленный срок.

Оформленная и подписанная студентом на титульном листе курсовая работа за одну неделю до защиты должна быть сдана научному руководителю. Несвоевременное представление курсовой работы приравнивается к неявке на экзамен, поэтому студент, не сдавший без уважительных причин в срок курсовую работу, получает неудовлетворительную оценку, считается имеющим академическую задолженность и не допускается к сдаче экзаменов. Преподаватель рассматривает курсовую работу в течение 3 дней с момента ее сдачи. Работа, получившая положительную оценку, допускается к защите и возвращается студенту для подготовки к защите. Отрицательная рецензия предполагает полную или частичную переработку курсовой работы, ее повторное рецензирование и (в случае положительной оценки) ее защиту. При подготовке к защите курсовой работы студент должен внимательно ознакомиться со всеми замечаниями, отмеченными руководителем, и устранить недостатки.

Защита курсовой работы – подведение итогов изучения дисциплины. Защита курсовой работы организуется руководителем курсовой работы и проводится в сроки, определенные расписанием занятий. Она проводится в присутствии студенческой группы, т.е. открыто. Защита студентами курсовых работ проводится перед комиссией в составе 2-3 преподавателей, назначаемых заведующим кафедрой. В состав комиссии включается и руководитель курсовой работы. Защита курсовой работы включает в себя следующие части:

1. презентация курсовой работы,
2. ответ на вопросы членов комиссии по теме работы.

Время доклада на защите курсовой работы не должно превышать 3-5 минут. Примерная структура доклада при защите курсовой работы приведена ниже:

– наименование темы работы;

– основные цели и задачи работы;

– логика построения работы в разделах работы;

– основные результаты, полученные в работе, выводы и обоснование предложений.

Защита курсовой работы имеет целью выявить глубину и самостоятельность знаний студента по избранной теме. На защите студент должен хорошо ориентироваться в представленной работе, уметь объяснить источники цифровых данных, отвечать на вопросы как теоретического, так и практического характера, относящиеся к теме работы.

Качество защиты курсовой работы оценивается на основании следующих показателей:

1. Форма доклада. Степень свободы и уверенности изложения материала, четкости мысли, корректности и правильности использования научно-технических понятий и терминов, лаконичность, умение использовать графический, иллюстративный материал.

2. Содержание доклада. Полнота, аргументированность и логическая последовательность изложения актуальности курсовой работы, ее цели и решаемых в ней задач, обоснование используемых методов решения, полученных в работе результатов, практических рекомендаций, выводов, доказательство их корректности, достоверности и практической значимости.

3. Адекватность восприятия. Степень адекватности восприятия, правильность и полнота ответов на поставленные вопросы.

4. Знание исходного теоретического материала. Знание студентом теоретических основ по теме курсовой работы.

5. Эрудированностъ. Уровень научной и инженерной эрудиции студента в теоретических и прикладных аспектах выполненной курсовой работы.

6. Уровень культуры. Внешний вид студента, культура поведения и речь, умение владеть собой.

2. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

При проведении опыта, целью которого является определение зависимости одной физической величины Y от другой физической величины X, неизбежно возникают ошибки измерения. Поэтому возникает задача: по имеющимся экспериментальным точкам (x_i, y_i) наилучшим образом воспроизвести искомую зависимость $Y(X)$. Для решения подобных задач часто применяют метод наименьших квадратов.

2.1. ОСНОВЫ ТЕОРИИ И РАСЧЕТНЫЕ ФОРМУЛЫ

Рассмотрим метод наименьших квадратов более подробно.

Пусть в результате эксперимента получена совокупность входных и выходных переменных в виде следующего набора:

x_{11}	x_{21}	…	x_{k1}	y_1
x_{12}	x_{22}	…	x_{k1}	y_2
.	.	…	.	.
.	.	…	.	.
.	.	…	.	.
x_{1N}	x_{2N}	…	x_{kN}	y_N

и задана функциональная зависимость в виде $y = \varphi(x_1, x_2, ... x_k)$.

Суть метода наименьших квадратов сводится к тому, чтобы сумма квадратов отклонений экспериментальных точек от сглаживающей кривой обращалась в минимум:

$$\sum_{i=1}^{N} [\, y_i - \varphi(x_i)]^2 = \min \qquad (2.1)$$

где y_i и x_i – экспериментальные значения переменных в i-том опыте, N – число опытов, $\varphi(x)$ – искомая зависимость y от x.

Известно, что через N точек можно всегда провести кривую, аналитически выражаемую многочленом $(N-1)$-й степени. Этот многочлен называют интерполяционным. А замену функции $y(x)$

на функцию $\varphi(x)$ так, что их значения совпадают в заданных точках $y(x_i)=\varphi(x_i);\ i=1,2,...,N$ называют интерполяцией.

Однако такое решение проблемы не является удовлетворительным, поскольку $y(x_i)=\varphi(x_i)$ из-за случайных ошибок измерения и влияния на измерения значений $y(x)$ различных помех.

Поэтому требуется провести кривую так, чтобы она в наименьшей степени зависела от случайных ошибок. Эта задача называется сглаживанием (аппроксимацией) экспериментальной зависимости и часто решается методом наименьших квадратов. Сглаживающую кривую называют аппроксимирующей.

При выборе вида зависимости $Y=f(X)$ возможны следующие случаи:

1. Общий вид зависимости $Y=f(X)$ известен заранее на основании теоретических предпосылок. Задача состоит в нахождении численных значений параметров этой зависимости.

2. Зависимость $Y=f(X)$ априори неизвестна, и нет никаких предположений о ее математической форме. В этом случае для эмпирического описания исследуемой закономерности в области ее существования, ограниченной пределами изменения аргумента, весьма удобно применить алгебраический полином определенной степени (ряд Тейлора):

$$Y=a_0+a_1X+x_2X^2+...+a_sX^s$$

Для определения параметров зависимости $Y=f(X)$ запишем ее как функцию не только аргумента X, но и параметров $a_j,\ j=\overline{0,\ s}$. Тогда условие (2.1) примет вид:

$$\sum_{i=1}^{N}\left[y_i-f\left(x_i,a_0,\ a_1,...,a_j,...a_s\right)\right]^2=\min \qquad (2.2)$$

Нахождение значений параметров, удовлетворяющих этому условию, сводится к отысканию экстремума функции многих переменных a_j. Для этого необходимо взять производные по параметрам от выражения (2.2) и приравнять их нулю, получив систему $s+1$ уравнений, решение которой дает возможность найти параметры a_j. Например при аппроксимации искомой зависимости полиномом s-того порядка $Y = a_0 + a_1 X + x_2 X^2 + ... + a_s X^s$ система уравнений примет вид:

$$
\begin{aligned}
&a_0 \cdot \sum_{i=1}^{N} x_i^0 + a_1 \cdot \sum_{i=1}^{N} x_i^1 + a_2 \cdot \sum_{i=1}^{N} x_i^2 + ... + a_s \cdot \sum_{i=1}^{N} x_i^s = \sum_{i=1}^{N} x_i^0 y_i \\
&a_0 \cdot \sum_{i=1}^{N} x_i^1 + a_1 \cdot \sum_{i=1}^{N} x_i^2 + a_2 \cdot \sum_{i=1}^{N} x_i^3 + ... + a_s \cdot \sum_{i=1}^{N} x_i^{s+1} = \sum_{i=1}^{N} x_i^1 y_i \\
&... \\
&a_0 \cdot \sum_{i=1}^{N} x_i^s + a_1 \cdot \sum_{i=1}^{N} x_i^{s+1} + a_2 \cdot \sum_{i=1}^{N} x_i^{s+2} + ... + a_s \cdot \sum_{i=1}^{N} x_i^{2s} = \sum_{i=1}^{N} x_i^s y_i
\end{aligned}
\tag{2.3}
$$

Эта система линейна относительно искомых параметров a_j и носит название нормальной. Ее решение дает численные значения параметров a_j.

Пусть изучается зависимость одной физической величины *Y* от другой *X*. В результате опытов получены *N* пар чисел $(x_i; y_i)$. При этом зависимость $y = \varphi(x)$ может быть представлена в виде

$$y(x) = a_0 + a_1 \cdot x$$

В соответствии с методом наименьших квадратов параметры этой зависимости должны удовлетворять условию

$$\sum_{i=1}^{N} \left[y_i - (a_0 + a_1 \cdot x_i) \right]^2 = \min \tag{2.4}$$

Взяв частные производные от выражения (2.4) по параметрам a_0 и a_1 и приравняв их нулю, получим систему нормальных уравнений:

$$\begin{aligned} a_0 \cdot \sum_{i=1}^{N} x_i^0 + a_1 \cdot \sum_{i=1}^{N} x_i &= \sum_{i=1}^{N} y_i \\ a_0 \cdot \sum_{i=1}^{N} x_i + a_1 \cdot \sum_{i=1}^{N} x_i^2 &= \sum_{i=1}^{N} y_i \cdot x_i \end{aligned} \qquad (2.5)$$

Решая систему уравнений (2.5), находим неизвестные параметры a_0 и a_1 и тем самым полностью определяем функцию, которая наилучшим образом (в смысле наименьших квадратов отклонений расчетных значений от фактических данных) аппроксимирует (приближает) искомую функцию $y(x)$.

Как видно из рассмотренного примера, изменение количества параметров не приведет к искажению сущности самого подхода, изменится лишь количество уравнений в системе (2.5) (для N параметров соответственно будет записано N уравнений).

После вычисления коэффициентов аппроксимирующего полинома необходимо провести проверку пригодности найденного полинома изучаемому объекту. Такую проверку будем называть проверкой адекватности аппроксимирующего полинома. То есть, необходимо проверить, насколько результат вычислений соответствует данным проведенного эксперимента, и является ли найденная зависимость соответствующей изучаемому явлению. Проверка адекватности заключается в доказательстве факта, что точность результатов, полученных по модели, будет не хуже точности расчетов, произведенных на основании экспериментальных данных.

Адекватность аппроксимирующей функции определяется при помощи коэффициента детерминации, найденного как отношение суммы квадратов регрессии к обей сумме квадратов:

$$R^2 = \frac{SSf}{SSf + SSr} \qquad (2.6)$$

где SSf – сумма квадратов регрессии, SSr – сумма квадратов остатков.

Сумма квадратов регрессии находится по формуле:

$$SSf = \sum_{i=0}^{N-1} \left(y(x_i) - Y_{cp} \right)^2 \qquad (2.7)$$

где $y(x_i)$ – расчетное значение функции отклика, Y_{cp} – среднее значение исходных данных функции отклика.

Сумма квадратов остатков находится по формуле:

$$SSr = \sum_{i=0}^{N-1} \left(Y_i - y(x_i) \right)^2 \qquad (2.8)$$

где Y_i – исходные значения функции отклика.

Общая сумма квадратов находится по формуле:

$$SS = \sum_{i=0}^{N-1} \left(Y_i - Y_{cp} \right)^2 \qquad (2.9)$$

При этом должно выполняться равенство:

$$SS = SSf + SSr \qquad (2.10)$$

Выражение (2.10) может служить проверочным расчетом при нахождении суммы квадратов регрессии и суммы квадратов остатков.

Таким образом, величина коэффициента детерминации R^2, лежащая в пределах $0 \le R^2 \le 1$ служит характеристикой адекватности математической модели объекту исследования.

Если коэффициент детерминации $R^2 < 0{,}7$, то найденная зависимость не адекватно описывает данные эксперимента. В этом случае необходимо выбрать другой вид зависимости.

Если $R^2 \ge 0{,}7$, то математическая модель признается адекватной экспериментальным данным.

2.2. ПОРЯДОК ВЫПОЛНЕНИЯ КУРСОВОЙ РАБОТЫ

Исходными данными для выполнения курсовой работы являются экспериментальные данные, представленные таблично.

1. Аппроксимировать экспериментальные данные линейной зависимостью вида $y_1(x) = a_0 + a_1 x$. Для этого составить систему нормальных уравнений (2.3) и решить ее относительно a_0 и a_1.

2. Построить на одном графике экспериментальные данные и найденную линейную зависимость $y_1(x) = a_0 + a_1 x$.

3. Выполнить проверку адекватности найденной зависимости экспериментальным данным с помощью коэффициента детерминации $R1^2$, найденного в соответствии с (2.6).

4. Аппроксимировать экспериментальные данные квадратичной зависимостью вида $y_2(x) = b_0 + b_1 x + b_2 x^2$. Для этого составить систему нормальных уравнений (2.3) и решить ее относительно b_0, b_1 и b_2.

5. Построить на одном графике экспериментальные данные и найденную квадратичную зависимость вида $y_2(x) = b_0 + b_1 x + b_2 x^2$.

6. Выполнить проверку адекватности найденной зависимости экспериментальным данным с помощью коэффициента детерминации $R2^2$, найденного в соответствии с (2.6).

7. Аппроксимировать экспериментальные данные экспоненциальной зависимостью вида $y_3(x) = a \cdot e^{bx}$.

8. Построить на одном графике экспериментальные данные и найденную экспоненциальную зависимость вида $y_3(x) = a \cdot e^{bx}$.

9. Выполнить проверку адекватности найденной зависимости экспериментальным данным с помощью коэффициента детерминации $R3^2$, найденного в соответствии с (2.6).

10. Сделать вывод о том, какая из полученных функций наилучшим образом аппроксимирует результаты эксперимента, и обосновать свои рассуждения.

3. ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА

Важнейшим условием научно поставленного эксперимента является минимизация общего числа опытов, следовательно, материальных и трудовых затрат и временных ресурсов, что, конечно, не должно существенно отражаться на качестве полученной информации.

Необходимо отметить еще одно важное обстоятельство. Применение методов планирования эксперимента ограничивается сложностью или невозможностью постановки экспериментов в реальных условиях. Однако методы моделирования позволяют проводить с помощью ЭВМ различные эксперименты с моделями объектов исследования. Принципиальных различий в планировании натурных экспериментов и исследований на ЭВМ нет, но в последнем случае больше возможностей изменять и стабилизировать любые переменные, которые учтены в модели. Функционирование реальной системы, связанное со случайным проявлением действия некоторых факторов, легко имитируется в машинных экспериментах с помощью специально разработанных программ.

3.1. ОСНОВЫ ТЕОРИИ И РАСЧЕТНЫЕ ФОРМУЛЫ

Планирование эксперимента состоит в выборе числа и условий проведения опытов, позволяющих получить необходимые знания об объекте исследования с требуемой точностью.

Рассмотрим сущность планирования эксперимента по этапам:

1. Выбор входных и выходных переменных. Входные переменные X_i, $i = \overline{1,k}$, которые определяют состояние объекта, назовем влияющими факторами. Основное требование к ним – достаточная управляемость, под которой понимается возможность установить нужный уровень фактора и стабилизировать его в течение всего опыта.

Выходная переменная Y – это реакция объекта на входные воздействия; она носит название функции отклика. Выбор этой функции определяется целью исследования, которая может представлять собой оптимизацию экономической (стоимость, производительность), технологической (точность, качество, быстродействие), конструктивной (габариты, надежность) или другой характеристики объекта.

2. Выбор области экспериментирования, т.е. области факторного пространства, изучение которой представляет интерес для исследования. Границы этой области по каждому фактору X_i обусловлены его минимальным и максимальным значениями, т.е. $X_{i\ \min} < X_i < X_{i\ \max}$. Область экспериментирования для случая учета двух факторов (X_1 и X_2) иллюстрируется рис. 3.1. При трех факторах такая область представляет собой параллелепипед. При большем числе факторов она ограничивается гиперплоскостями в k-мерном пространстве.

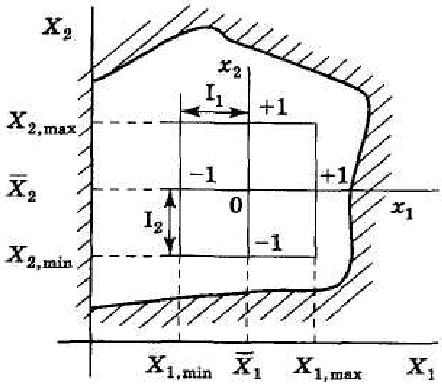

Рис. 3.1. Область экспериментирования для двух факторов

Оценка границ области определения (существования) факторов производится на основе принципиальных ограничений либо из других соображений.

Первый вид ограничений не может быть нарушен ни при каких обстоятельствах. Например, для температуры нижним пределом всегда будет абсолютный нуль.

При выборе ограничений второго вида исследователь руководствуется конкретными обстоятельствами, например: временем протекания процесса, стоимостью материала и т.п. Устанавливая область определения, необходимо выполнять также условие совместимости факторов, т.е. значения факторов должны быть выбраны так, чтобы эксперимент можно было реализовать.

3. Выбор математической модели объекта. Если аналитическую зависимость, связывающую функцию отклика Y с влияющими факторами X_i, найти невозможно и вид функции $Y = f(X_1, X_2, ..., X_k)$ априори неизвестен, то целесообразно использовать степенной ряд

$$Y = a_0 + \sum_{i=1}^{k} a_i X_i + \sum_{i<j} a_{ij} X_i X_j + \sum_{i=1}^{k} a_{ii} X_i^2 + \dots \ , \qquad (3.1)$$

где k – число влияющих факторов.

Выражение (3.1) служит математической моделью исследуемого объекта. Так как, исходя из требований практики, число членов степенного ряда ограничивается, аппроксимирующая функция представляет собой полином некоторой степени.

Для определения коэффициентов аппроксимирующего полинома применяется наиболее универсальный метод наименьших квадратов. Как отмечалось выше, при его использовании необходимым условием получения статистических оценок является выполнение неравенства $N > s$, т.е. количество опытов N должно быть больше, чем число коэффициентов полинома s. Увеличить N можно повторением опыта в исходных точках эксперимента либо увеличением количества этих точек. Второй путь дает возможность не только учесть погрешности измерения, но и оценить адекватность аппроксимирующего полинома во всей области экспериментирования.

Для удобства обработки и интерпретации результатов эксперимента целесообразно все факторы представить в безразмерной форме, для чего производят операцию кодирования переменных. Ее сущность заключается в том, что начало координат факторного пространства переносится в точку с координатами $\overline{X_i}$ (рис. 3.1, центр эксперимента – точка 0), где $\overline{X_i} = 0{,}5(X_{i\ \max} + X_{i\ \min})$. Кроме того, интервал варьирования факторов $I_i = 0{,}5(X_{i\ \max} - X_{i\ \min})$ разбивается на ряд уровней, симметричных относительно центра эксперимента. В случае составления симметричных двухуровневых планов все k

факторов изменяются на двух уровнях, при этом значениям $X_{i\,\max}$ отвечает кодированная переменная $x_i = +1$, а значениям $X_{i\,\min}$ соответствует $x_i = -1$. Для количественных факторов связь между физическими (X_i) и кодированными (x_i) значениями факторов определяется соотношением $x_i = \left(X_i - \overline{X_i}\right)/I_i$.

Описанные преобразования являются линейными. Поэтому в аппроксимирующей функции (3.1) изменяются только коэффициенты при факторах, т.е.

$$Y = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i<j} b_{ji} x_i x_j + \sum_{i=1}^{k} b_{ji} x_i^2 + \ldots , \qquad (3.2)$$

где x_i – влияющие факторы в безразмерной форме (независимые переменные).

Заметим, что число членов полинома (3.2) при практической аппроксимации обычно ограничивается учетом линейного и квадратичного влияния факторов, а также эффекта парного их взаимодействия.

4. Составление плана эксперимента. Выбрав математическую модель объекта, определяют, какое значение должен принимать каждый из факторов в каждом из опытов. Таблица, составленная из значений факторов для каждого опыта, как независимых ($i = \overline{1, k}$) так и зависимых ($i = \overline{k+1,\ s-1}$), называется матрицей планирования, ее часть, которая включает в себя значения независимых переменных, и является планом эксперимента.

Эксперимент, в котором реализуются все возможные сочетания уровней факторов, называется полным факторным экспериментом (ПФЭ).

Если k факторов варьируются на двух уровнях, то число всех возможных сочетаний факторов равно 2^k. Тогда полный факторный эксперимент называется ПФЭ типа 2^k. Если число уровней факторов составляет n, необходим ПФЭ типа n^k.

В качестве примера рассмотрим составление плана ПФЭ 2^2. Число опытов в этом случае $N = 2^2 = 4$. Соответствующая матрица планирования представлена таблице 3.1.

Таблица 3.1

Номер опыта	x_0	x_1	x_2	$x_3 = x_1 \cdot x_2$	Y
1	+1	–1	–1	+1	Y_1
2	+1	+1	–1	–1	Y_2
3	+1	–1	+1	–1	Y_3
4	+1	+1	+1	+1	Y_4

Данный план соответствует модели вида

$$Y = b_0 x_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2. \qquad (3.3)$$

Первый столбец матрицы занимает фиктивный фактор $x_0 = +1$ при коэффициенте полинома b_0. Столбцы матрицы x_1 и x_2, обведенные рамкой (табл. 3.1), задают планирование: по ним непосредственно определяются условия опытов; столбец x_3 является самостоятельным и заполняется по данным столбцов x_1 и x_2. Он, как и столбец x_0, используется при расчетах.

По результатам эксперимента, проведенного в соответствии с представленным планом, можно определить все четыре коэффициента полинома (3.3). Однако в этом случае $s = 4$ и, следовательно, условие $N > s$ не выполняется, что не позволяет произвести статистических оценок аппроксимирующей зависимости. Для получения таких оценок нужно ограничиться линейной зависимостью без учета взаимовлияния факторов (при этом $s = 3 < N = 4$) либо провести дополнительный опыт в нулевой точке $x_1 = x_2 = 0$ (тогда $s = 4 < N = 5$).

Если ошибка опыта априори неизвестна, для ее оценки опыты в каждой точке плана нужно повторить.

В общем случае ПФЭ типа 2^k обладает следующими свойствами:

– *симметричностью относительно центра эксперимента*; при этом алгебраическая сумма элементов вектора-столбца для каждого фактора равна нулю:

$$\sum_{u=1}^{N} x_{iu} = 0,\ i = \overline{1,k}, \tag{3.4}$$

где x_{iu} – значение i-го фактора в u-м опыте;

– *соответствием условиям нормировки* (сумма квадратов элементов каждого столбца равна числу опытов):

$$\sum_{u=1}^{N} x_{iu}^2 = N,\ i = \overline{0,k} \tag{3.5}$$

Это условие следует из того обстоятельства, что значения факторов в матрице задаются в кодированном виде $(-1;+1)$;

– *условиям ортогональности*, при этом сумма почленных произведений любых двух векторов-столбцов матрицы равна нулю:

$$\sum_{u=1}^{N} x_{iu} x_{ju} = 0,\ i < j;\ i, j = \overline{1,k} \tag{3.6}$$

Ортогональность матриц планирования позволяет при обработке данных с помощью метода наименьших квадратов получить независимые друг от друга оценки коэффициентов.

При выполнении условий (3.4) ÷ (3.6) коэффициенты аппроксимирующего полинома рассчитываются по формулам:

$$b_i = \frac{1}{N}\sum_{u=1}^{N} Y_u\, x_{iu};\ b_{ij} = \frac{1}{N}\sum_{u=1}^{N} Y_u\, x_{iu} x_{ju} \tag{3.7}$$

Заметим, что модель типа (3.3) включает только линейные эффекты и эффекты парного взаимодействия факторов. Планы, соответствующие этой модели, называются планами первого порядка. В случае необходимости учета нелинейного влияния фактора аппроксимирующий полином должен содержать члены более высокого порядка. В частности, модель с квадратичными членами для двух факторов (полная квадрика) имеет вид

$$Y = b_0 x_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2. \tag{3.8}$$

Для оценки коэффициентов полинома вида (3.8) следует пользоваться более сложными планами, чем планы первого порядка. Это обусловлено тем обстоятельством, что помимо информации о коэффициентах при линейных членах и членах, учитывающих эффект взаимодействия, необходимо оценить коэффициенты b_i при квадратичных членах аппроксимирующего полинома. Планы этого типа называются планами второго порядка.

3.2. ПОРЯДОК ВЫПОЛНЕНИЯ КУРСОВОЙ РАБОТЫ

Для обработки результатов эксперимента необходимо воспользоваться следующей схемой:

1. На основании данных параллельных наблюдений оценить дисперсию воспроизводимости D_{y_i} для каждой строки плана по формуле:

$$D_{y_i} = \frac{1}{m-1} \sum \left(y_{ij} - \overline{y_i} \right)^2 \tag{3.9}$$

где $\overline{y_i} = \frac{1}{m} \sum_{j=1}^{m} y_{ij}$ – среднее значение y в i-том опыте (незакрашеный кружок на рис. 3.2), y_{ij} – экспериментальные значения функции отклика (черные точки на рис. 3.2), m – число повторений i-того опыта.

2. Определить критерий Кохрена G, который представляет собой отношение максимальной дисперсии воспроизводимости к сумме всех дисперсий, по формуле:

$$G = \frac{D_{y\ \max}}{\sum D_{y_i}} \tag{3.10}$$

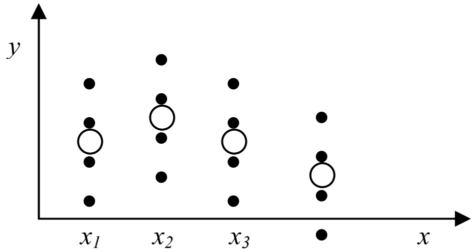

Рис. 3.2. Графическая интерпретация результатов эксперимента:

После этого выполнить проверку однородности дисперсий (погрешности опыта). Для этого рассчитанное по (3.10) значение критерия Кохрена G при принятом уровне значимости α (α – вероятность отвергнуть правильную гипотезу при условии, что она верна, чаще всего $\alpha = 0{,}05$) сравнить с табличным значением G_T (прил. 1), которое является функцией числа степеней свободы $m-1$ и N.

Если $G < G_T$, то гипотеза о равноточности дисперсий воспроизводимости не отвергается. При этом признается, что все опыты выполнены с равной погрешностью.

Если условие $G < G_T$ не выполняется, то делают вывод, что опыты выполнены не с равной погрешностью, и эксперимент необходимо переделать.

3. Найти коэффициенты аппроксимирующего полинома (3.3) по формулам (3.7). Записать уравнение полинома и найти расчетные значения функции отклика.

4. Проверить гипотезу об адекватности модели изучаемому объекту по критерию Фишера *F*, найденного по формуле:

$$F = \frac{D_{y_a}}{D_{y_o}} \tag{3.11}$$

Здесь D_{y_a} – дисперсия адекватности, характеризующая рассеивание экспериментальных значений *y* относительно аппроксимирующей кривой, определяемая по формуле:

$$D_{y_a} = \frac{1}{N-s} \sum_{i=1}^{N} \left(y_{p\,i} - y_{э\,i} \right)^2 \qquad (3.12)$$

где N – число опытов, s – число коэффициентов полинома, $y_{p\,i}$ и $y_{эi}$ – расчетные и экспериментальные значения y в i-том опыте соответственно.

D_{y_o} – дисперсия опыта, характеризующая рассеяние значений выходного параметра y в эксперименте, и рассчитывается по формуле:

$$D_{y_o} = \frac{1}{mN} \sum D_{y_i} \qquad (3.13)$$

где D_{y_i} – дисперсия воспроизводимости опытов, найденная по формуле (3.9), mN – общее число опытов.

Для оценки качества регрессионной модели выполняется сравнение рассчитанного по (3.11) значения критерия Фишера F и табличного значения F_T (прил. 2), зависящим от числа степеней свободы $N(m-1)$ и $N-s$.

Табличное значение критерия Фишера F_T – это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости.

Если $F < F_T$, то гипотеза об адекватности аппроксимирующей зависимости экспериментальным данным не отвергается (математическая модель признается адекватной объекту исследования).

Если неравенство $F < F_T$ не удовлетворяется, то гипотеза об адекватности аппроксимирующей зависимости экспериментальным данным отвергается (математическая модель признается неадекватной объекту исследования). В данном случае следовало бы либо увеличить число опытов эксперимента, либо поменять вид аппроксимирующей зависимости.

5. Проверить значимость коэффициентов полинома (3.3) по критерию Стьюдента (отдельно по каждому коэффициенту), рассчитанному по формуле:

$$t_{b_i} = \frac{|b_i|}{\sigma_{b_i}} \qquad (3.14)$$

где $|b_i|$ – абсолютное значение коэффициента аппроксимирующего полинома; σ_{b_i} – среднее квадратическое отклонение коэффициентов аппроксимирующего полинома b_i, $\sigma_{b_i} = \sqrt{D_{b_i}}$; D_{b_i} – дисперсия коэффициентов аппроксимирующего полинома b_i, $D_{b_i} = \frac{D_{y_o}}{N}$.

Значение t_{b_i} сравнивается с табличным t_T (прил. 3). Если $t_{b_i} < t_T$, то коэффициент b_i считается не значимым (т.е. можно принять $b_i = 0$) и соответствующее слагаемое исключается из уравнения регрессии. Если $t_{b_i} > t_T$, то коэффициент b_i считается значимым (существенно влияющим на функцию отклика).

После сравнения расчетных значений t_{b_i} и табличным t_T, из рассмотрения исключаются незначимые коэффициенты, и осуществляется повторная проверка адекватности модели.

Таким образом, оценка значимости коэффициентов аппроксимирующей зависимости позволяет исключить из уравнения незначимые коэффициенты, тем саамы упростить вид аппроксимирующей зависимости.

6. Завершить вычисления интерпретацией модели в терминах объекта исследования. Прежде всего, выяснить, в какой мере каждый из факторов влияет на функцию отклика. Значения линейных коэффициентов служат количественной мерой, оценивающей влияние факторов: чем больше коэффициент b_i, тем сильнее это влияние. Знак коэффициента позволяет судить о характере зависимости функции отклика от соответствующих факторов.

Затем аналогично следует проанализировать эффект парных взаимодействий.

4. ВЫРАВНИВАНИЕ СТАТИСТИЧЕСКИХ РЯДОВ

Законом распределения случайной величины X называется любое правило, позволяющее находить вероятность P всевозможных событий, связанных с этой случайной величиной. Например, вероятность того, что случайная величина X не примет какое-то значение или попадет в определенный интервал. В последнем случае обычно пользуются так называемой плотностью распределения непрерывной случайной величины $f(x)$. При этом вероятность попадания этой величины в интервал $a < x < b$ будет

$$P\{a < x < b\} = \int_a^b f(x)dx \tag{4.1}$$

а интеграл от плотности распределения в пределах $(-\infty; +\infty)$ равен единице.

4.1. ОСНОВЫ ТЕОРИИ И РАСЧЕТНЫЕ ФОРМУЛЫ

Предположим, на основе эксперимента имеем n наблюдений случайной величины X, которые разделим на k интервалов Δx_i. Подсчитаем количество значений этой величины m_i, приходящихся на каждый интервал. Тогда частота появления X в интервале Δx_i равна

$$p_i^* = \frac{m_i^*}{n} \tag{4.2}$$

Составим статистический ряд (табл. 4.1), в котором приведены интервалы в порядке их расположения по оси X, а также соответствующие им частоты.

Число интервалов, в которые, следует группировать статистический материал, не должно быть слишком большим (тогда ряд распределения становится невыразительным, и частоты в нем обнаруживают незакономерные колебания); с другой стороны, оно не должно быть слишком малым (при малом числе разрядов свойства распределения описываются статистическим рядом слишком грубо).

Практика показывает, что в большинстве случаев рационально выбирать порядка 10-20 интервалов.

Таблица 4.1

Статистический ряд

Δx_i	$[x_0, x_1]$	$[x_1, x_2]$	…	$[x_i, x_{i+1}]$	…	$[x_{k-1}, x_k]$
p_i^*	p_1^*	p_2^*	…	p_{i+1}^*	…	p_k^*

Статистический ряд часто оформляется графически в виде *гистограмм.* Для построения гистограммы по оси абсцисс откладывают интервалы Δx_i и на каждом из них, как на основании, строят прямоугольник, площадь которого равна p_i^* (рис. 4.1). Таким образом, из условия построения следует, что площадь гистограммы равна единице.

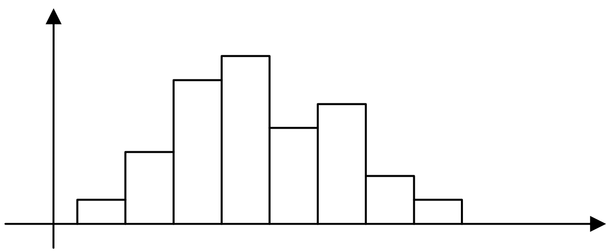

Рис. 4.1. Гистограмма распределения случайной величины

Имея гистограмму распределения, можно построить статистическую функцию распределения.

С помощью гистограммы распределения можно подобрать наиболее вероятную теоретическую кривую плотности распределения случайной величины *X*. Эта задача называется выравниванием статистических распределений. Наилучшее приближение статистической зависимости к имеющемуся статистическому материалу выбирается, исходя из соображений, связанных с физикой решаемой задачи или с внешним видом гистограммы.

Пользуясь данными статистического ряда, можно приближенно построить и статистическую функцию распределения $F^*(x)$ величины *X*. Для практики обычно достаточно построить статистическую функцию распределения по нескольким точкам. В качестве этих точек удобно взять границы интервалов $x_0, x_1, ..., x_k$, которые фигурируют в статистическом ряде. Тогда,

$$
\begin{aligned}
&F^*(x_0) = 0;\\
&F^*(x_1) = p_1^*;\\
&F^*(x_2) = p_1^* + p_2^*;\\
&\ldots\\
&F^*(x_i) = \sum_{i=0}^{i} p_i^*;\\
&\ldots\\
&F^*(x_k) = \sum_{i=0}^{k} p_i^* = 1.
\end{aligned}
\tag{4.3}
$$

Соединяя полученные точки ломаной линией или плавной кривой, получим приближенный график статистической функции распределения $F^*(x)$. Пример гистограммы распределения и статистической функции распределения приведены на рис. 4.2 и 4.3 соответственно.

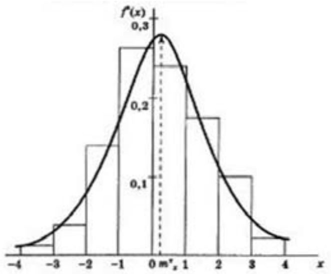

Рис. 4.2. Гистограмма распределения величины X

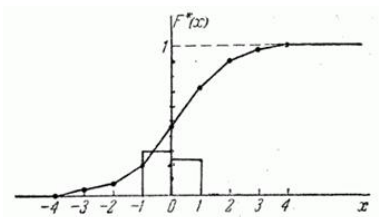

Рис. 4.3. Статистическая функция распределения

Во многих случаях при обработке экспериментальных данных предполагается, что распределение случайных величин подчиняется *нормальному закону*, или закону ошибок, введенному Гауссом. Следует отметить, что в большинстве случаев это предположение оправдано. Однако если априори закон распределения или его параметры неизвестны, то следует проверить правильность соответствующих гипотез.

Непрерывная случайная величина X называется распределенной по *нормальному закону*, если ее плотность распределения имеет вид:

$$f(x) = \frac{1}{\sigma_x \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{(x - m_x)^2}{2\sigma_x^2} \right) \tag{4.4}$$

где m_x и σ_x – математическое ожидание и среднее квадратическое отклонение случайной величины X соответственно.

Нормализованным нормальным распределением называется такое нормальное распределение, у которого $m_x = 0$ и $\sigma_x = 1$. Из нормализованного распределения можно получить любое другое нормальное распределение с заданными m_x и σ_x по формуле: $z = m_x + x \cdot \sigma_x$.

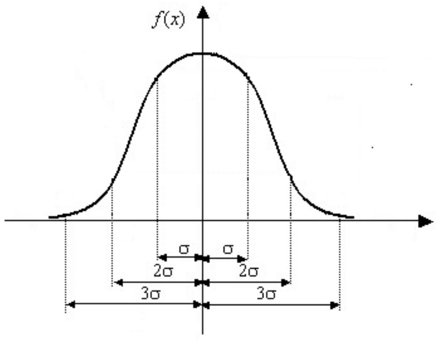

Рис. 4.4. Графический вид нормального закона распределения случайной величины X с параметрами $m_x = 0$ и $\sigma_x = 1$ (распределение нормализовано)

График плотности нормального распределения имеет вид колокола. На рис. 4.4 показано нормализованное нормальное распределение.

График (рис. 4.4) показывает, что в области $-\sigma_x < x < \sigma_x$ на графике сосредоточено 68 % площади распределения, в области

$-2\sigma_x < x < 2\sigma_x$ – сосредоточено 95,4 % площади распределения, в области $-3\sigma_x < x < 3\sigma_x$ – сосредоточено 99,7 % площади распределения. Таким образом, практически все значения нормально распределенной случайной величины лежат в интервале $\left[m_x - 3\sigma_x ; m_x + 3\sigma_x\right]$. Это правило называется «правилом трех сигм». Другими словами, если отклонение случайной величины X превосходит $3\sigma_x$, то это свидетельствует либо о грубой ошибке в опыте (так называемый «промах»), либо о том, что случайная величина распределена не в соответствии с нормальным законом распределения. И в том, и в другом случаях объем эксперимента следует увеличить.

Свойства нормального распределения

Как видно из (4.4), нормальное распределение имеет два параметра: математическое ожидание m_x и среднеквадратичное отклонение σ_x величины X от этого математического ожидания m_x.

Изменение параметра нормального распределения m_x приводит к сдвигу кривой по оси x (рис. 4.5).

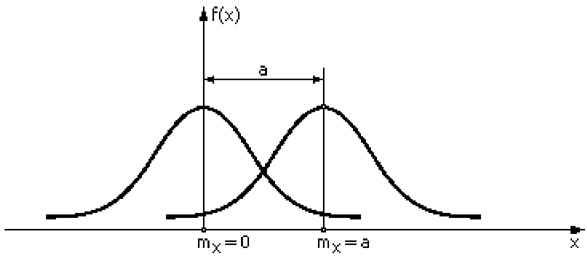

Рис. 4.5. Влияние параметра «математическое ожидание» на вид закона нормального распределения случайной величины X

Изменение параметра нормального распределения σ_x приводит к масштабированию формы по оси x (рис. 4.6). При этом в любом случае площадь под кривой плотности вероятности всегда неизменна и равна 1.

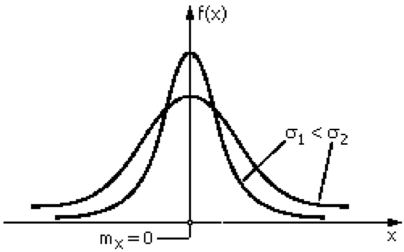

Рис. 4.6. Влияние параметра «среднее квадратическое отклонение» на вид закона нормального распределения случайной величины X

Рассмотрим подробнее функцию распределения случайной величины, распределенной по нормального закону. Данная функция носит название функции Лапласа (рис. 4.7) и задается интегралом от плотности вероятности нормального распределения:

$$F(x)=\frac{1}{\sigma_x \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{(x-m_x)^2}{2\sigma_x^2}\right) dx \qquad (4.5)$$

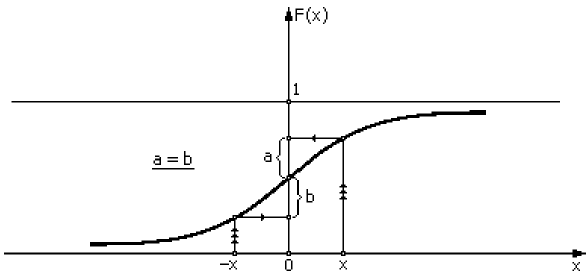

Рис. 4.7. Функция Лапласа

Этот интеграл не берется в общем виде, поэтому функция Лапласа задана в виде таблицы (прил. 4) для $m_x = 0$ и $\sigma_x = 1$.

Поскольку функция Лапласа является симметричной относительно точки $(0; 0{,}5)$, то $F(-x) = 1 - F(x)$. Поэтому в таблице содержится только одна из ее симметричных частей.

Если задается интервал интегрирования функции Лапласа $[a, b]$, то:

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

$$= \int_{\frac{a-m_x}{\sigma_x}}^{\frac{b-m_x}{\sigma_x}} \frac{1}{\sigma_x \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{t^2}{2} \right) dt = \Phi\left(\frac{b - m_x}{\sigma_x} \right) - \Phi\left(\frac{a - m_x}{\sigma_x} \right) \tag{4.6}$$

Выравнивание статистических рядов

Во всяком статистическом распределении неизбежно присутствуют элементы случайности, связанные с тем, что число наблюдений ограничено, что произведены именно те, а не другие опыты, давшие именно те, а не другие результаты. Только при очень большом числе наблюдений эти элементы случайности сглаживаются, и случайное явление обнаруживает в полной мере присущую ему закономерность. На практике мы почти никогда не имеем дела с таким большим числом наблюдений и вынуждены считаться с тем, что любому статистическому распределению свойственны в большей или меньшей мере черты случайности. Поэтому при обработке статистического материала часто приходится решать вопрос о том, как подобрать для данного статистического ряда теоретическую кривую распределения, выражающую лишь существенные черты статистического материала, но не случайности, связанные с недостаточным объемом экспериментальных данных. Такая задача называется задачей выравнивания (сглаживания) статистических рядов.

Задача выравнивания заключается в том, чтобы подобрать теоретическую плавную кривую распределения, с той или иной точки зрения наилучшим образом описывающую данное статистическое распределение (рис. 4.8).

Предположим, что исследуемая величина $X-$ это ошибка измерения, возникающая в результате суммирования воздействий множества независимых элементарных ошибок, тогда из теоретических соображений можно считать, что величина X подчиняется нормальному закону (4.4) и задача выравнивания переходит в задачу о рациональном выборе параметров m_x и σ_x.

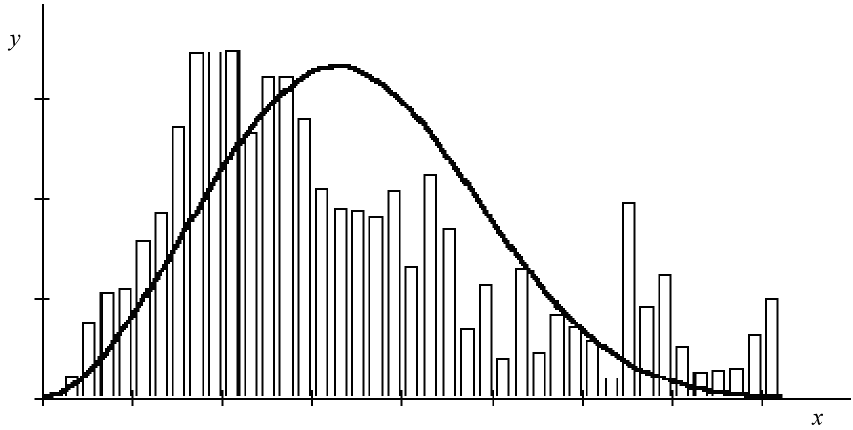

Рис. 4.8 Гистограмма и теоретическая кривая

Таким образом, требуется подобрать параметры m_x и σ_x так, чтобы функция $f(x)$ наилучшим образом описывала данный статистический материал.

Один из методов, применяемых для решения этой задачи, является метод моментов. Согласно методу моментов, параметры m_x и σ_x выбираются с таким расчетом, чтобы несколько важнейших числовых характеристик (моментов) теоретического распределения были равны соответствующим статистическим характеристикам. То есть, должны выполняться условия $m_x = m_x^*$ и $\sigma_x = \sigma_x^*$.

Статистическое математическое ожидание находится по формуле:

$$m_x^* = \sum_{i=1}^{k} m_{x_i} \cdot p_i^* \tag{4.7}$$

где m_{x_i} – среднее значение интервала, p_i^* – частота появления величины появления X в i-том интервале.

Статистическое среднее квадратическое отклонение находится в соответствии с выражением:

$$\sigma_x^* = \sqrt{D_x^*} \tag{4.8}$$

где D_x^* – дисперсия величины *X,* определяемая по формуле:

$$D_x^* = \alpha^* - \left(m_x^*\right)^2 \tag{4.9}$$

где α – второй начальный момент, определяемый выражением:

$$\alpha^* = \sum_{i=1}^{k} m_{x_i}^2 \cdot p_i^* \tag{4.10}$$

Критерий согласия Пирсона

Предположим, что гистограмма может быть аппроксимирована аналитической функцией $f(x)$. Как бы хорошо ни была подобрана теоретическая кривая, между нею и статистическим распределением неизбежны некоторые расхождения.

Естественно возникает вопрос: объясняются ли эти расхождения только случайными обстоятельствами, связанными с ограниченный числом наблюдений, или они являются существенными и связаны с тем, что подобранная нами кривая плохо выравнивает данное статистическое распределение. Для ответа на этот вопрос может быть использован <u>критерий согласия</u> χ^2 (читается «хи-квадрат»), предложенный Пирсоном.

Предположим, что произведено n независимых опытов, в каждом из которых случайная величина X приняла определенное значение. Результаты опытов сведены в статистический ряд (табл. 4.1). Требуется проверить, согласуются ли экспериментальные данные с гипотезой о том, что случайная величина X имеет данный закон распределения (заданный функцией распределения $F(x)$ или плотностью распределения $f(x)$). Назовем этот закон распределения «теоретическим».

Зная закон распределения, можно найти теоретические вероятности попадания случайной величины в каждый из интервалов: p_1, p_2,..., p_k.

Проверяя согласованность теоретического и статистического распределений, будем исходить из расхождений между теоретическими вероятностями p_i и наблюденными частотами p_i^*.

В качестве меры расхождения между теоретическим и статистическим распределениями выбрана сумма квадратов отклонений между вероятностями, взятых с некоторыми «весами» c_i:

$$\chi^2 = \sum_{i=1}^{k} c_i \left(p_i^* - p_i\right)^2 \tag{4.11}$$

Коэффициенты c_i («веса» интервалов) вводятся потому, что в общем случае отклонения, относящиеся к различным разрядам, нельзя считать равноправными по значимости. Действительно, одно и то же по абсолютной величине отклонение $\left(p_i^* - p_i\right)$, может быть мало значимым, если сама вероятность p_i мала. Поэтому «веса» c_i необходимо взять обратно пропорциональными вероятностям p_i попадания случайной величины в каждый из интервалов.

К. Пирсон показал, что если положить

$$c_i = \frac{n}{p_i} \tag{4.12}$$

то при большом числе опытов n закон распределения величины χ^2 обладает простыми свойствами: он практически не зависит от функции распределения $F(x)$ и от числа опытов n, а именно, этот закон при увеличении n приближается к так называемому «распределению χ^2».

Таким образом, с учетом (4.12) выражение (4.11) примет вид:

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i} \left(p_i^* - p_i\right)^2 \tag{4.13}$$

Учитывая, что $p_i^* = \dfrac{m_i}{n}$, где m_i – число значений в i-том интервале, получим:

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i} \qquad (4.14)$$

Распределение χ^2 зависит от параметра r, называемого числом «степеней свободы» распределения. Число «степеней свободы» r равно числу интервалов k за вычетом числа независимых условий («связей»), наложенных на частоты p_i^*. Обычно на p_i^* накладывают следующие условия s:

1. Сумма частот p_i^* должна равняться единице.

$$\sum_{i=1}^{k} p_i^* = 1$$

2. Средние значения теоретического и статистического распределений должны совпадать.

$$\sum_{i=1}^{k} x_i \cdot p_i^* = m_{x_i}$$

3. Должны совпадать теоретическая и статистическая дисперсии.

$$\sum_{i=1}^{k} \left(x_i - m_{x_i}\right)^2 \cdot p_i^* = D_x$$

Для распределения χ^2 составлены таблицы (прил. 5). Пользуясь этими таблицами, можно для каждого значения χ^2 и числа степеней свободы r найти вероятность p того, что величина, распределенная по закону χ^2, превзойдет это значение.

Если найденное значение p велико, то гипотезу о том, что функция $f(x)$ пригодна для аппроксимации полученной гистограммы, можно признать не противоречащей опытным данным. Если же вероятность p мала (менее 0,1), то эта гипотеза отбрасывается как неправдоподобная.

4.2. ПОРЯДОК ВЫПОЛНЕНИЯ КУРСОВОЙ РАБОТЫ

Индивидуальное задание содержит n выборочных данных.

1. Для составления статистического ряда необходимо разбить данные на 6-10 равных интервалов. Для этого найти x_{min} и x_{max}. Записать размах выборки $\Delta = x_{\max} - x_{\min}$. Определить длину интервала $h = \Delta / k$, где k – число интервалов (выбрать самостоятельно от 6 до 10 интервалов).

2. Составить статистический ряд (табл. 4.2).

Таблица 4.2

Δx_i	$[x_0, x_1]$	$[x_1, x_2]$	…	$[x_i, x_{i+1}]$	…	$[x_{k-1}, x_k]$
m_i	m_1	m_2	…	m_{i+1}	…	m_k
p_i^*	p_1^*	p_2^*	…	p_{i+1}^*	…	p_k^*

Записать последовательно интервалы Δx_i. Найти количество значений m_i, приходящихся на каждый интервал. Подсчитать для каждого интервала частоту p_i^* попадания данных в каждый интервал. Убедиться, что сумма всех частот p_i^* равна 1.

3. Построить гистограмму распределения. Для этого на оси *OX* отложить все интервалы и на каждом из них построить прямоугольник высотой p_i^* / h.

4. На этом же графике построить статистическую плотность распределения. Для этого соединить отрезками точки с координатами $\left(\overline{x_i}; p_i^* / h \right)$, где $\overline{x_i}$ - средние значение *i*-того интервала.

5. Выдвинуть гипотезу о нормальном распределении результатов эксперимента. Для этого параметры m_x и σ_x принять такими, чтобы математическое ожидание m_x и среднее квадратическое отклонение σ_x теоретического распределения совпадали с соответствующими статистическими характеристиками m_x^* и σ_x^*. То есть, должны выполняться условия $m_x = m_x^*$ и $\sigma_x = \sigma_x^*$.

6. Рассчитать числовые значения параметров m_x^* и σ_x^*.

7. Записать теоретическую функцию плотности распределения $f(x)$. Найти ее значения для середин отрезков i-того интервала. $\overline{x_i}$. Записать результаты вычислений в таблицу 4.2, добавив соответствующую строку.

8. Построить на одном графике гистограмму распределения, графики статистической и теоретической плотности распределения.

9. Вычислить меру расхождения χ^2 между теоретическим и статистическим распределениями по критерию Пирсона.

10. Определить число степеней свободы $r = k - s$.

11. По r и χ^2 с помощью таблиц (прил. 5) определить вероятность того, что величина, имеющая распределение χ^2 с r степенями свободы, превзойдет данное значение χ^2. Если эта вероятность весьма мала (менее 0,1), то гипотеза отбрасывается как неправдоподобная. В этом случае следует либо выбрать другой вид распределение, либо провести дополнительные эксперименты.

Если эта вероятность относительно велика (больше 0,1), то гипотезу можно признать не противоречащей опытным данным. В этом случае признается, что случайная величина, представленная выборкой, имеет нормальный закон распределения.

ПРИЛОЖЕНИЯ

Приложение 1

Табличное значение критерия Кохрена

N	$m-1$								
	1	**2**	**3**	**4**	**5**	**7**	**9**	**16**	**36**
2	0,998	0,975	0,939	0,906	0,877	0,833	0,801	0,734	0,660
3	0,967	0,871	0,798	0,746	0,707	0,653	0,617	0,547	0,475
4	0,906	0,768	0,684	0,629	0,589	0,536	0,502	0,437	0,372
5	0,841	0,684	0,598	0,544	0,506	0,456	0,424	0,364	0,307
6	0,781	0,616	0,532	0,480	0,445	0,398	0,368	0,314	0,261
7	0,727	0,561	0,480	0,431	0,397	0,354	0,326	0,276	0,228
8	0,680	0,516	0,438	0,391	0,360	0,318	0,293	0,246	0,202
9	0,638	0,478	0,403	0,358	0,329	0,290	0,266	0,223	0,182
10	0,602	0,445	0,373	0,331	0,303	0,267	0,244	0,203	0,166
15	0,471	0,335	0,276	0,242	0,220	0,191	0,174	0,143	0,114
20	0,389	0,270	0,220	0,192	0,174	0,150	0,136	0,111	0,088
30	0,293	0,198	0,159	0,138	0,124	0,106	0,096	0,077	0,060
60	0,174	0,113	0,090	0,076	0,068	0,058	0,052	0,041	0,032

Приложение 2

Табличное значение критерия Фишера

$N(m-1)$	$N-s$								
	1	**2**	**3**	**4**	**5**	**6**	**8**	**10**	**15**
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,40	19,43
3	10,13	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,70
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,86
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,62
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	3,94
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64	3,51
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35	3,22
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,14	3,01
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,98	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,85	2,75	2,62
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,60	2,46
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,49	2,35
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,41	2,27
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,35	2,20

Приложение 3

Табличное значение критерия Стьюдента

$N(m-1)$	t_T	$N(m-1)$	t_T
1	12,7060	11	2,2010
2	4,3020	12	2,1788
3	3,1820	13	2,1604
4	2,7760	14	2,1448
5	2,5700	15	2,1314
6	2,4460	16	2,1190
7	2,3646	17	2,1098
8	2,3060	18	2,1009
9	2,2622	19	2,0930
10	2,2281	20	2,0860

Приложение 4

Табличное значение функции Лапласа

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,00000	1,00	0,34134	2,00	0,47725	3,00	0,49865	4,00	0,49997
0,05	0,01994	1,05	0,35314	2,05	0,47982	3,05	0,49886	4,05	0,49997
0,10	0,03983	1,10	0,36433	2,10	0,48213	3,10	0,49903	4,10	0,49998
0,15	0,05962	1,15	0,37493	2,15	0,48422	3,15	0,49919	4,15	0,49998
0,20	0,07926	1,20	0,38493	2,20	0,48610	3,20	0,49932	4,20	0,49999
0,25	0,09871	1,25	0,39435	2,25	0,48778	3,25	0,49928	4,25	0,49999
0,30	0,11791	1,30	0,40320	2,30	0,48928	3,30	0,49952	4,30	0,49999
0,35	0,13683	1,35	0,41149	2,35	0,49062	3,35	0,49960	4,35	0,49999
0,40	0,15542	1,40	0,41924	2,40	0,49180	3,40	0,49967	4,40	0,49999
0,45	0,17364	1,45	0,42647	2,45	0,49286	3,45	0,49972	4,45	0,50000
0,50	0,19146	1,50	0,43319	2,50	0,49379	3,50	0,49977	4,50	0,50000
0,55	0,20884	1,55	0,43943	2,55	0,49462	3,55	0,49981	4,55	0,50000
0,60	0,22575	1,60	0,44520	2,60	0,49534	3,60	0,49984	4,60	0,50000
0,65	0,24215	1,65	0,45053	2,65	0,49598	3,65	0,49987	4,65	0,50000
0,70	0,25804	1,70	0,45543	2,70	0,49654	3,70	0,49989	4,70	0,50000
0,75	0,27337	1,75	0,45994	2,75	0,49702	3,75	0,49991	4,75	0,50000
0,80	0,28814	1,80	0,46407	2,80	0,49745	3,80	0,49993	4,80	0,50000
0,85	0,30234	1,85	0,46784	2,85	0,49782	3,85	0,49994	4,85	0,50000
0,90	0,31594	1,90	0,47128	2,90	0,49814	3,90	0,49995	4,90	0,50000
0,95	0,32894	1,95	0,47441	2,95	0,49841	3,95	0,49996	4,95	0,50000

Приложение 5

Табличное значение распределения χ^2

p / r	**0,99**	**0,95**	**0,90**	**0,80**	**0,50**	**0,20**	**0,10**	**0,05**	**0,01**
1	0,00	0,01	0,02	0,06	0,45	1,64	2,71	8,84	6,64
3	0,12	0,35	0,58	1,01	2,37	4,64	6,26	7,82	11,8
5	0,65	1,14	1,61	2,84	4,35	7,29	9,24	11,1	15,1
7	1,24	2,17	2,38	3,82	6,35	9,80	12,0	14,1	18,5
9	2,09	3,32	4,17	5,38	8,34	12,2	14,7	16,9	21,7
12	3,57	5,23	6,30	7,81	11,3	16,8	18,5	21,0	28,2
16	5,23	7,26	8,55	10,8	14,8	19,8	22,8	25,0	80,6
18	7,02	9,39	10,9	12,9	17,8	22,8	26,0	28,9	34,8

СПИСОК РЕКОМЕНДОВАННОЙ ЛИТЕРАТУРЫ

1. Власов, Константин Петрович. Методы научных исследований и организации эксперимента: Учеб. пособие / Ред. А.А. Гальнбек; С.-Петерб. гос. горн. ин-т. Каф. печей, контроля и автоматизации металлургического производства. - СПб: СПГГИ, 2000

2. Горелов, С.В. Основы научных исследований: учебное пособие / С.В. Горелов, В.П. Горелов, Е.А. Григорьев; под ред. В.П. Горелова. - 2-е изд., стер. - Москва; Берлин: Директ-Медиа, 2016 http://biblioclub.ru/index.php?page=book_view_red&book_id=443846

3. Диков, А.В. Математическое моделирование и численные методы: учебное пособие / А.В. Диков, С.В. Степанова ; под ред. Г.В. Сугробова. - Пенза : ПГПУ, 2000 http://biblioclub.ru/index.php?page=book_view_red&book_id=96973

4. Клинов, А.В. Математическое моделирование химико-технологических процессов: учебное пособие / А.В. Клинов, А.Г. Мухаметзянова; Федеральное агентство по образованию, Государственное образовательное учреждение высшего профессионального образования "Казанский государственный технологический университет". - Казань: Казанский государственный технологический университет, 2009 http://biblioclub.ru/index.php?page=book_view_red&book_id=270540

4. Планирование и организация эксперимента [Электронный ресурс]: практикум / Новосиб. гос. аграр. ун-т. Биолого-технолог. фак; сост. И.А. Ленивкина. – Новосибирск, 2012 http://znanium.com/catalog.php?bookinfo=516007

5. Планирование научного эксперимента: Учебник/В.А.Волосухин, А.И.Тищенко, 2-е изд. - М.: ИЦ РИОР, НИЦ ИНФРА-М, 2016 http://znanium.com/bookread2.php?book=516516

6. Статистические методы обработки экспериментальных данных с использованием пакета MathCad: Учебное пособие/Ф.И.Карманов, В.А.Острейковский - М.: КУРС, НИЦ ИНФРА-М, 2015 http://znanium.com/bookread2.php?book=508241

ОГЛАВЛЕНИЕ

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Методические указания к курсовой работе для студентов бакалавриата направления 15.03.04

Сост. *Н.В. Васильева*

Печатается с оригинал-макета, подготовленного кафедрой автоматизации технологических процессов и производств

Ответственный за выпуск *Н.В. Васильева*

Лицензия ИД № 06517 от 09.01.2002

Подписано к печати 26.06.2019. Формат 60×84/16.
Усл. печ. л. 2,6. Усл.кр.-отт. 2,6. Уч.-изд.л. 2,3. Тираж 100 экз. Заказ 606. С 218.

Санкт-Петербургский горный университет
РИЦ Санкт-Петербургского горного университета
Адрес университета и РИЦ: 199106 Санкт-Петербург, 21-я линия, 2