УЧЕБНО-НАУЧНЫЙ КОМПЛЕКС «ИНФОРМАТИКА» ТГУ

В. П. ЛЕОНОВ

ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ НА ПРОГРАММИРУЕМЫХ МИКРОКАЛЬКУЛЯТОРАХ

(ПРИКЛАДНАЯ СТАТИСТИКА НА Б3-34, МК-52, МК-54, МК-56, МК-61)

Практическое пособие

Под редакцией Б. А. Гладких

ИЗДАТЕЛЬСТВО ТОМСКОГО УНИВЕРСИТЕТА
Томск — 1990

УДК 681.321+519.2(076)

Леонов В. П. **Обработка экспериментальных данных на программируемых микрокалькуляторах (Прикладная статистика на Б3-34, МК-52, МК-54, МК-56, МК-61).** — Томск: Изд-во Том. ун-та, 1990. — 376 с. — 3 р. 20 к. 5000 экз. 1502000000.

В монографии изложены методы прикладной статистики для обработки экспериментальных данных и реализующие их 60 программ. Кроме методов оценки основных статистических характеристик — среднего, дисперсии, и т. д. и корреляционно-регрессионного анализа, дано изложение однофакторного и двухфакторного дисперсионного анализа, а также ковариационного анализа. Несколько глав посвящены методам проверки гипотез о векторах средних, дискриминантному анализу и методу главных компонент.

Использование методов и программ иллюстрируется более чем 100 задачами с решениями из области биологии, медицины, ботаники, геологии, полупроводниковой электроники, сельского хозяйства, экономики, социологии, гидрометеорологии, химии, психологии и т. д.

Для студентов и специалистов, желающих научиться эффективному применению ПМК в обработке экспериментальных данных.

Рецензенты — В. С. Трусов, Г. А. Иващенко

ISBN 5—7511—0254—1

Л $\frac{1502000000}{177(012)-90}$ 37—87

© Леонов В. П., 1990

Василий Петрович ЛЕОНОВ

ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ НА ПРОГРАММИРУЕМЫХ МИКРОКАЛЬКУЛЯТОРАХ (Прикладная статистика на Б3-34, МК-52, МК-54, МК-56, МК-61)

Практическое пособие

Редактор **Л. И. Дюканова.**
Художественный редактор **Ю. Г. Григорьев.**
Технический редактор **Г. Н. Гридина.**
Корректор **О. П. Наумова.**
ИБ 1863.

Сдано в набор 24.11.88 г. Подписано к печати 24.01. 90 г. КЗ09002. Формат 84×108 1/32. Бумага типографская № 3. Литературная гарнитура. Высокая печать. Печ. л. 11,75. Усл. печ. л. 19,74. Уч.-изд. л. 20,2. Тираж 5000 экз. Заказ 341. Цена в перепл. 3 р. 20 к.

Издательство ТГУ, 634029, Томск, ул. Никитина,4.
Тип. «Красн. рабочий», г. Красноярск, пр. Мира, 91.

ПРЕДИСЛОВИЕ

...знание людей заслуживает имени Науки в зависимости от того, какую роль в нем играет число.

Э. Борель

Применение методов прикладной статистики в научных исследованиях сопряжено с необходимостью проведения большого объема вычислений, что, в свою очередь, требует использования ЭВМ. Стационарные ЭВМ, типа СМ или ЕС ЭВМ, стоимость часа машинного времени которых составляет от нескольких десятков до сотен и тысяч рублей, требуют двух- или трехсменной загрузки. Только в этом случае возможна рентабельная эксплуатация этих ЭВМ в течение их срока службы. Вследствие этого стационарные ЭВМ по сути дела являются мощными вычислительными средствами коллективного пользования. По этой причине режим работы стационарных ЭВМ таков, что время ожидания результата решения отдельных задач, даже если чистое машинное время для их решения секунды или минуты, составляет часы или дни. С другой стороны, портативность, малая стоимость, простота пользования и доступность программируемых микрокалькуляторов (ПМК) в сочетании с развитием их программного обеспечения все больше способствуют превращению их в мощные «личные инструменты» для выполнения вычислительных задач различного характера.

Использование ПМК в качестве персональных микроЭВМ характеризуется тремя отличительными особенностями.

1. При решении достаточно большого класса задач использование ПМК гораздо оперативнее, эффективнее и экономически более целесообразно, нежели использование больших ЭВМ. От 50 до 90% задач обработки экспериментальных данных (малых и средних объ-

емов), ежедневно возникающих в практике исследователя и требующих оперативного применения методов прикладной статистики, вполне могут быть решены с использованием ПМК. Как правило, легче и лучше научить специалиста в конкретной области и математике, и программированию, чем наоборот. Это утверждение полностью относится и к обучению работе на ПМК.

2. Персональное использование ПМК для вычислений, оперативность проверки готовых программ на тестовых примерах и возможность быстрой корректировки этих программ самим пользователем способствует значительному расширению числа активно применяемых исследователем методов вычислений и овладению новыми, ранее недоступными методами.

3. Язык программирования этих калькуляторов весьма прост и нагляден. В то же время этот язык содержит и такие команды, как команды условных переходов, команда безусловного перехода, есть и команды организации циклов (аналогичные командам IF, GOTO и DO). Это позволяет достаточно быстро реализовывать тот или иной алгоритм в виде соответствующей программы. Человека, владеющего навыками работы на ПМК и умеющего программировать на ПМК, значительно легче научить более сложным языкам программирования, например, Бэйсику или Фортрану.

За последние годы изданы справочники Я. К. Трохименко и Ф. Д. Любича, А. Н. Цветкова и В. А. Епанечникова, В. П. Дьяконова, содержащие программы по разнообразным методам вычислений на ПМК. Однако развитие программного обеспечения ПМК делает необходимым выпуск и более узкоспециализированных изданий.

Указанные выше обстоятельства делают весьма актуальной задачу издания таких пособий, в которых бы доступно излагались: 1) методы прикладной статистики по обработке экспериментальных данных; 2) их программные реализации для ПМК; 3) контрольные решения типовой задачи для каждой программы; 4) набор задач с их решениями для этих программ по различным отраслям науки и техники. Предлагаемая читателю книга содержит эти четыре компоненты и, будучи ориентирована прежде всего на практиков, преследует следующие цели:

1. Дать представление о возможностях ПМК и научить читателя вводить и контролировать ввод программ в ПМК.

2. Познакомить в доступной форме с идеями и основами математического аппарата ряда методов прикладной статистики.

3. Дать программные реализации этих методов и инструкции к этим программам.

4. Развить у читателя устойчивый навык решения на ПМК задач обработки экспериментальных данных с помощью изложенных методов, используя для этой цели около 100 контрольных примеров и задач с их решениями из различных областей исследований.

5. Подготовить читателя к знакомству с более глубокими по своему содержанию книгами по методам прикладной статистики, часть из которых приведена в списке литературы.

Навык достаточно свободной ориентации в методах, описанных в этой книге, зависит, естественно, от уровня подготовки исследователя и его практического опыта применения этих и подобных методов. Для начинающих овладение данными методами можно порекомендовать следующий порядок работы с книгой. Первоначально изучить идеи метода и его математический аппарат. Затем научиться правильному вводу программы, реализующей изученный метод, и добиться решения контрольного примера, используя инструкцию к программе. После этого решить задачи к этой главе и сравнить полученные решения с решениями, приведенными в конце главы. Наконец, следующий, очень важный этап состоит в том, чтобы, используя приведенные в каждой главе задачи, попытаться сгенерировать условия новых задач, исходя из конкретной области исследований каждого читателя. Именно такая тренировка по наполнению содержательным смыслом «голых» цифровых данных этих задач позволяет перейти к следующему этапу: постановке и решению конкретных задач, представляющих самостоятельную практическую и научную ценность в конкретной области исследований читателя. По мнению автора, такая последовательность позволит читателю подняться до уровня понимания используемых методов.

Тематика примеров и задач для иллюстрации возможностей рассматриваемых методов обработки и реализующих их программ достаточно широка. В частности, использованы примеры и задачи по технике, технологии, экономике, геологии, биологии, химии, медицине, экспериментальной ботанике, сельскому хозяйству, метеорологии и т. д.

Подавляющее большинство примеров и задач для их решения на ПМК специально взято из достаточно известных специалистам пособий и монографий. Это позволяет читателю при желании познакомиться по первоисточнику более глубоко с содержательной интерпретацией постановки решаемой задачи и полученного решения. Однако не для всех задач одинаково подробно приводятся результаты решения этих задач с помощью тех или иных программ. Для ряда задач в их решениях приводится указание идентифицировать приведенные числа как те или иные выборочные параметры, что можно сделать только после самостоятельного решения задач на ПМК.

Все программы с инструкциями вынесены в отдельную часть — библиотеку программ. Это облегчает использование книги как пособия для начинающих изучение методов прикладной статистики и ее применение как библиотеки программ для ПМК по статистическому анализу экспериментов. Большое внимание уделено непосредственно расчетам, которые изложены с такой подробностью, чтобы можно было полностью овладеть подчас достаточно сложной системой вычислений.

Автор выражает свою благодарность директору УНВК «Информатика» при ТГУ Б. А. Гладких за ряд ценных замечаний по структуре книги и содержанию примеров и задач.

ГЛАВА 1

РАЗВИТИЕ ПРИКЛАДНОЙ СТАТИСТИКИ И ЕЕ РОЛЬ В ОБРАБОТКЕ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Некоторые люди ненавидят само имя статистика, но я нахожу ее исполненной красоты и интереса.

Ф. Гальтон

1.1. От государствоведения до ... статистики

Пожалуй ни одна из наук не претерпела столь удивительных превращений на пути своего развития, как статистика. Немногим известен тот факт, что первоначально, примерно два с половиной столетия тому назад, статистика определялась как государствоведение (наука об управлении). В коллекции высказываний о термине «статистика» есть, например, и такое определение: «Статистика есть наука, которая учит нас тому, каково политическое устройство всех современных государств известного нам мира» [1]. В 1842 г. Британская энциклопедия давала такое определение этому термину: «Слово «статистика» происходит от слова «государство» (state или Staat) и означает группу людей, живущих в общественном союзе; оно включает все характеристики их состояния» [1]. И только примерно сто лет назад, удовлетворяя все возрастающие потребности развивающихся научных исследований и промышленных преобразований, статистика начинает свое превращение в инструмент исследования массовых явлений, в том числе и обработки экспериментальных наблюдений. Интенсивное развитие статистики происходило в период первой и второй мировых войн. Это вызывалось как повышением требований к качеству военной продукции, так и тем, что в промышленности в эти периоды увеличивалось число неквалифицированной рабочей силы, заменившей призванные в армию квалифицированные кадры. Это приводило к необходимости совершенствования и развития статистических методов контроля качества продукции, а также методов статистического регулирования технологических процессов.

Развитие статистики как науки привело к появлению отдельных, специальных ее направлений, например, статистики финансов, статистики труда, статистики материально-технического снабжения и т. д. В отдельную научную дисциплину выделилась и прикладная статистика.

1.2. Предмет прикладной статистики

Современное определение прикладной статистики дается как «...самостоятельной научной дисциплины, разрабатывающей и систематизирующей понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации и обработки (в том числе с помощью ЭВМ) статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов» [2, с. 19].

До некоторых пор применение методов прикладной статистики для обработки экспериментальных данных было в основном прерогативой представителей технических наук. Однако в последние десятилетия это положение стало стремительно меняться. Интенсивное использование прикладной статистики наблюдается в последнее время в экономике, геологии, медицине, биологии, сельском хозяйстве, иммунологии, политологии, психологии, педагогике и многих других науках.

Все большее число ранее описательных экспериментальных наук становятся точными науками. Пожалуй, наиболее типично в этом плане развитие социально-экономических наук [4, 5]. Математическая статистика сейчас претендует на роль метаязыка экспериментатора.

Применение методов прикладной статистики необходимо в тех случаях, когда число экспериментов ограничено и в результатах этих экспериментов наблюдается невоспроизводимость. Причины подобной неидентичности этих результатов кроются в наличии неконтролируемых и нерегулируемых факторов, влияющих на эксперимент. К этим же последствиям приводит и различие объектов экспериментов (отличающиеся по качеству исходные материалы и т. п.), невозможность абсолютно

точного воспроизведения всего комплекса контролируемых и регулируемых факторов эксперимента.

Все это в совокупности и приводит к рассеянию количественных результатов экспериментов. Такое рассеяние и невоспроизводимость результатов создает ситуацию неопределенности в принятии оптимальных решений на основе этих результатов. Именно эту особенность и подчеркивает одно из современных определений статистики, которое было дано Абрахамом Вальдом: статистика — совокупность методов, которые дают нам возможность принимать оптимальные решения в условиях неопределенности.

Сегодня для всех направлений научных исследований (от биологии до ядерной физики) прикладная статистика — азбука науки экспериментальных исследований. В совокупности с навыками программирования и применения ЭВМ она составляет ту самую вторую грамотность, которой должен владеть сегодня каждый современный исследователь. Будучи инструментом экспериментальных исследований методы прикладной статистики позволяют резко сокращать объемы этих исследований и более точно, определенно формулировать их результаты.

ГЛАВА 2

РАСЧЕТЫ НА ПРОГРАММИРУЕМЫХ МИКРОКАЛЬКУЛЯТОРАХ

Цель расчетов — не числа,
а понимание.
Г. В. Хемминг

Программируемые микрокалькуляторы (ПМК) индивидуального пользования могут работать в двух режимах: в режиме калькулятора и в режиме автоматических вычислений. В первом случае пользователь осуществляет ввод данных и команд на выполнение необходимых операций вручную. В том случае, когда последовательность выполнения этих операций известна и работа в ручном режиме занимает длительное время, можно внести эту последовательность в специальное запоминающее устройство ПМК, называемое программной памятью. После ввода программы, которая и представляет собой последовательность выполняемых операций, в программную память ПМК может выполнить ее в автоматическом режиме за более короткий отрезок времени.

2.1. Редактирование программы

Очень часто при вводе программы в режиме программирования могут быть допущены ошибки, являющиеся результатом невнимательности. Возможно также недостаточное нажатие на клавишу, при котором не происходит замыкания электрических контактов. Даже опытные пользователи, постоянно работающие с ПМК, допускают ошибки при вводе программ. Поэтому нужно взять за правило проверку правильности введенной в ПМК программы, тем более, что такая проверка занимает не более 1,5÷2 мин, а у опытных пользователей и меньше, даже если длина программ максимальна. Редактирование программы после ее ввода осуществляется путем возврата программы к началу, т. е. обычно к

адресу 00 с помощью клавиши В/О, затем вновь переход в режим программирования F ПРГ. При последовательном нажатии клавиши $\overrightarrow{\text{ШГ}}$ производится просмотр кодов команд и их сравнение с кодами команд, приведенных в таблице программы. При выполнении такой процедуры не надо забывать, что в ПМК индицируется адрес не последней команды, а следующей за ней, т. е. адрес последней команды меньше на 1, чем индицируемый адрес. В том случае, когда в программе обнаружена лишняя команда, ее можно заменить операцией (нет операции), К НОП. Если обнаружена ошибочно введенная команда, то следует нажатием клавиши $\overleftarrow{\text{ШГ}}$ сместить программу на шаг назад, так что справа на индикаторе индицируется адрес именно этой команды, и далее нажать соответствующие клавиши.

Часто при составлении программ на ПМК 34 либо их корректировке требуется изменить адрес перехода (безусловного, условного либо косвенного). В этом случае необходимо вернуться на 2 шага назад, т. е. вновь набрать команду перехода, а затем и адрес перехода. При нарушении этого требования в программную память записывается неправильный адрес.

При работе с вновь набранной программой необходимо взять за правило даже после проверки правильности ввода программы решение контрольного примера. Это не занимает много времени, т. к. обычно авторы программ в качестве контрольного примера используют малую выборку. Однако при подтверждении результатов по контрольному примеру нет сомнений в правильности полученных результатов по собственным введенным данным. При разборе новой программы и выяснении алгоритма ее работы целесообразно проводить ее выполнение в пошаговом режиме, нажимая для этого клавишу ПП. При каждом нажатии на эту клавишу происходит выполнение одной только команды программы и остановка ПМК. Опыт разработки и разбора программ для ПМК позволяет также рекомендовать фиксацию содержимого всех операционных регистров X, Y, Z и T ПМК 34 в отдельной таблице по каждому шагу. Такая запись содержимого регистров способствует лучшему пониманию работы программы, а также поиску ее резервов.

При выполнении некорректных операций (например, извлечение квадратного корня из отрицательного числа, деление на 0, вычисление логарифма отрицательного числа и т. п.) счет на ПМК прекращается, а на индикаторе появляется специальный знак ошибки.

В ПМК 34 при всех подобных ошибках на индикаторе появляется специальный знак ЕГГОГ, представляющий собой искаженное английское слово ERROR — ошибка.

2.2. Анализ и изучение особенностей работы программы

Для овладения умением самостоятельного программирования на ПМК начинающему пользователю можно посоветовать следующее. Внимательно изучить руководство по эксплуатации имеющейся ПМК. Достаточно полное понимание возможностей ПМК, как правило, приходит после нескольких чтений руководства. Изучить также имеющуюся по данному вопросу литературу [6—11]. Эффективно проведение анализа используемых программ с записью содержимого операционных регистров на каждом шаге программы и сопоставлением этих результатов с командами, в результате которых получен очередной результат. Без таких упражнений невозможно овладеть умением быстро и грамотно составлять необходимые программы либо разбираться в чужих программах. Можно рекомендовать также материалами по ПМК, регулярно публикуемые в журнале «Наука и жизнь» в разделе «Человек с микрокалькулятором», а также в журналах «Техника — молодежи», «Химия и жизнь», «Квант» и др.

В качестве примера предлагаемого анализа используемых программ проанализируем программу построения гистограммы распределения псевдослучайной величины, определяемой как дробная часть выражения $(\pi+x_i)^5$ [11]. Программа с указанием адресов команд, самих команд и их кодов приведена ниже и предназначена для ПМК 34. Такая форма представления и записи программ, используемая в литературе, представляется нам наиболее удачной и полной, т. к. позволяет после ввода программы произвести быстрый контроль пра-

вильности ввода согласно адресу команды и ее коду. Именно в форме таких таблиц и оформлены далее все приводимые программы. Кроме самой программы для ее правильного использования дается «Инструкция к программе», в которой сообщается о порядке ввода данных, о промежуточных и окончательных результатах. Ниже приводится такая инструкция к разбираемой программе. Значение числа экспериментов (псевдослучайных чисел) возьмем равным 100, т. е. $n=100$ (табл. 1, 2).

Таблица 1

Адрес	Команда	Код	Адрес	Команда	Код	Адрес	Команда	Код
00	5	05	09	ИПД	6Г	18	КИПД	ГГ
01	ИПС	6[	10	—	11	19	1	01
02	Fπ	20	11	ПС	4[	20	+	10
03	+	10	12	1	01	21	КПД	LГ
04	FX^Y	24	13	0	00	22	F L0	5Г
05	↑	ОЕ	14	X	12	23	00	00
06	ПД	4Г	15	1	01	24	С/П	50
07	КИПД	ГГ	16	+	10			
08	F'*	25	17	ПД	4Г			

* Здесь и далее символом F' будем обозначать команду вращения операционного стека.

Для более детального рассмотрения особенностей построения программы и ее работы оформим в виде таблицы результаты, хранящиеся на каждом шаге программы в операционных регистрах ПМК 34 с указанием команды и адреса команды. В табл. 3 первые четыре строки соответствуют регистрам X, Y, Z и T, а в пятой строке указана команда и через черту ее адрес в программе. В регистрах указываются результаты, которые получаются после выполнения соответствующей команды. К началу программы в регистрах Y, Z и T имеется 0, а в регистре X число $n=100$, оставшееся при занесении n в регистр 0 согласно п. 3 инструкции. После нажатия клавиши ПП на индикаторе появляется число 5 — показатель степени в выражении $(\pi+x_i)^5$, которое введено в регистр X по команде 5, расположенной

Таблица 2

№ п/п	Содержание	Набрать	Выполнить	Результат
1.	Ввести программу	F ПРГ	ввод FАВТ	
2.	Ввести 0 в регистры 1÷9, А, С	0	П1 П2 П3 П4 П5 П6 П7 П8 П9 ПА ПС	0
3.	Ввести в регистр 0 число экспериментов	n	ПО	n
4.	Начать вычислительный эксперимент по построению гистограммы распределения дробной части выражения $(\pi+x_1)^5$ в интервале 0÷1 (с разбиением на 10 подынтервалов)		В/О С/П	n_1 в П1 n_2 в П2 n_3 в П3 n_4 в П4 n_5 в П5 n_6 в П6 n_7 в П7 n_8 в П8 n_9 в П9 n_{10} в ПА

по адресу 00. Вновь нажимаем клавишу ПП — на индикаторе 0. Это число появилось в результате команды ИПС, расположенной в программе по адресу 01, и представляет собой значение X_1. После очередного нажатия ПП на индикаторе появляется число $\pi=$ $=3{,}1415926$ — это результат выполнения команды по адресу 02. Вновь нажимаем ПП, при этом выполняется операция «+» по адресу 03, т. е. происходит сложение X_1 и π, в результате вновь индицируется число 3,1415926, т. к. $X_1=0$. Обратите при этом внимание на то, как происходит перемещение чисел из одного операционного регистра в другой при выполнении тех или иных команд. На следующем шаге программы происходит выполнение операции возведения числа 3,1415926 в пятую степень, т. е. выполняется команда Fx^Y, где $Y=5$. В результате этого на индикаторе появляется число $306{,}01951=$ $=(3{,}1415926)^5$. По следующей команде ↑ это число из регистра X вводится в регистр Y, а затем по команде ПД производится засылка этого числа в регистр памяти Д. Следующие за этим 4 шага программы с адреса 07 по адрес 10 служат для того, чтобы выделить дробную часть полученного при выполнении команды Fx^Y

Таблица 3

X	100	5	0	3,141 592 6	3,141 592 6
Y	0	100	5	0	5
Z	0	0	100	5	100
T	0	0	0	100	100
	Команды →	5/00	ИПС/01	Fπ/02	+/03
X	306, 019 51	306,019 51	306,019 51	0	306,019 51
Y	5	306,019 51	306,019 51	306,019 51	306,019 51
Z	100	5	5	306,019 51	5
T	100	100	100	5	0
	Fx^Y/04	↑/05	ПД/06	КИПД/07	FΘ/08
X	000 003 06	$1{,}951 \cdot 10^{-2}$	$1{,}951 \cdot 10^{-2}$	1	10
Y	306,019 51	306,019 51	306,019 51	$1{,}951 \cdot 10^{-2}$	$1{,}951 \cdot 10^{-2}$
Z	306,019 51	5	5	306,019 51	306,019 51
T	5	5	5	5	5
	ИПД/09	—/10	ПС/11	1/12	0/13
X	$1{,}951 \cdot 10^{-1}$	1	1,195,1	1,195 1	0
Y	306,019 51	$1{,}951 \cdot 10^{-1}$	306,019 51	306,019 51	1,195 1
Z	5	306,019 51	5	5	306,019 51
T	5	5	5	5	5
	×/14	1/15	+/16	ПД/17	КИПД/18
X	1	1	1	1	1
Y	0	1,195 1	1,195 1	1,195 1	1,195 1
Z	1,195 1	306,019 51	306,019 51	306,019 51	306,019 51
T	306,019 51	306,019 51	306,019 51	306,019 51	306,019 51
	1/19	+/20	КПД/21	FL0/22	00/23
X	5	$1{,}951 \cdot 10^{-2}$	3,141 592 6	3,161 102 6	315,640 47
Y	1	5	$1{,}951 \cdot 10^{-2}$	5	5
Z	1,195 1	1	5	1	1
T	306,019 51	1,195,1	1	1	1
	5/00	ИПС/01	Fπ/02	+/03	Fx^Y/04
X	315,640 47	315,640 47	1	315,640 47	0000 003 15
Y	315,640 47	315,640 47	315,640 47	315,640 47	315,640 47
Z	5	5	315,640 47	5	315,640 47
T	1	1	5	1	5
	↑ /05	ПД/06	КИПД/07	FΘ/08	ИПД/09

числа 306,01951. Для этого используется следующее свойство косвенного обращения к регистрам памяти. Если в регистр памяти (в нашем случае это регистр памяти Д) помещено дробное число больше +1, то после выполнения операции косвенного обращения к этому регистру памяти в нем остается только целая часть этого числа. Последующий вызов этой целой части числа и вычитание его из исходного числа и позволяет выделить дробную часть этого числа в интервале от 0 до 1. Следовательно, по команде КИПД, расположенной по адресу 07, произойдет выделение целой части числа 306,01951 в регистре Д. Какое при этом вызовется случайное число на индикатор и из какого регистра, для наших вычислений значения не имеет. По следующей команде F' мы производим вращение операционного стека, благодаря чему это случайное число помещается в регистр Т, а затем и вовсе удаляется из операционного стека. Но благодаря этой команде F' из регистра Y перемещается число 306,01951. Следующей командой ИПД (адрес 09) мы вызываем в регистр X целую часть этого числа из регистра памяти Д, а по команде — (адрес 10) находим уже и дробную часть этого числа, равную $1{,}951 \cdot 10^{-2}$, которую и помещаем в регистр памяти С. Сохранение этого числа в регистре памяти С необходимо, так как согласно программе по адресу 01 происходит вызов последнего значения x_i. Следовательно, если $X_1=0$, то $X_2=1{,}951 \cdot 10^{-2}$. По командам 1 и 0, расположенным по адресам 12 и 13, в операционном регистре X производится набор чисел 10, после чего по команде $\times$ (адрес 14) производится умножение дробной части $1{,}951 \cdot 10^{-2}$ на число 10. По команде 1 (адрес 15) в регистре X набирается число 1, с которым далее и складывается предыдущий результат, полученный при умножении. Эта операция сложения выполняется командой + по адресу 16. Полученный при этом результат 1,1951 по команде ПД (адрес 17) помещается в регистр памяти Д. Именно эта и последующие 4 команды по 21-й адрес включительно и производят непосредственно размещение полученного числа ($1{,}951 \cdot 10^{-2}$) в один из счетчиков, в роли которых выступают регистры с 1 по 9 и регистр А (10). По команде КИПД (адрес 18) происходит косвенное обращение к одному из регистров памяти. Как уже можно догадаться, при выполнении команды

КИПД произойдет обращение к регистру 1, так как в регистре Д помещено число $1<1{,}1951<2$. При выполнении команды КИПД учитывается только целая часть хранимого в регистре Д числа, которая и равна 1. Поскольку перед пуском программы в регистры памяти $1\div9$ и А (10) занесено число 0, то после выполнения команды КИПД из регистра 1 вызовется в операционный регистр число 0. По команде 1 (адрес 19) в регистре X наберется число 1, при этом число 0 переместится в регистр Y, далее эти два числа складываются по команде + (адрес 20) и результат заносится в тот же самый первый регистр по команде КПД. Команды FL0 (адрес 22) и 00 (адрес 23) обеспечивают возврат на адрес 00 после выполнения одного цикла. Если после этого обратиться в регистр 0 командой ИПО, то на индикаторе появится число 00000099, т. е. осталось еще 99 циклов до окончания заданного объема вычислений, после которого выполнится команда С/П (адрес 24), а поскольку 100 циклов еще не выполнено, то второй цикл начинается вновь набором числа 5, затем из регистра С командой ИПС (адрес 01) вызывается число $X_2=1{,}951\cdot10^{-2}$ и т. д. Если после набора программы выполнить инструкцию к ней и, нажав клавиши В/О С/П, запустить ПМК в автоматический режим счета, то через $17\div18$ минут ПМК прекратит вычисления, выполнив заданные 100 циклов. На индикаторе при этом будет число 17. Для того чтобы определить, сколько чисел, представляющих собой дробные части выражения $(\pi+x_i)^5$ и расположенных на отрезке $0\div1$, попали в промежутки $0\div0{,}1$; $0{,}1\div0{,}2$; ... $0{,}9\div1$, поочередно выполним команды обращения к регистрам памяти $1\div9$, А (10), нажимая для этого клавиши ИП1, ИП2 ... ИПА. В результате этого будет индицироваться последовательный ряд чисел: 17; 4; 9; 15; 14; 8; 6; 9; 14; 4. Сумма этих чисел равна 100. Выбрав соответствующий масштаб, нетрудно построить гистограмму. Несложно и преобразовать исходное выражение. Так, если мы захотим проверить возможность аппроксимации равномерного распределения аналогичным выражением $(\pi+x_i)^7$, то в этом случае по адресу 00 вместо 5 набирается 7. При этом 5002 сгенерированных по этому алгоритму числа, точнее их дробные части, разместятся в 10 интервалах следующим образом: 514; 896; 768; 259; 512; 258; 512;

258; 637; 388. Счет при этом занимает уже несколько часов. Из приведенных результатов видно, что полученное распределение далеко от равномерного. Если же по адресу 00 вместо 5 или 7 набрать команду $F\pi$, то мы получим иной алгоритм генерации $(\pi+x_i)^{\pi}$, 3000 наблюдений по этому алгоритму имеют следующее распределение: 310; 266; 313; 328; 253; 280; 265; 418; 299; 268. Видно, что за исключением интервала $0{,}7 \div 0{,}8$, на котором число показаний (частота) равна 418, остальные интервалы имеют более равномерное распределение, чем вышеприведенный пример.

Внимательно разобрав приведенную выше программу, попытайтесь на ее основе сконструировать свою программу для построения данных, вводимых извне, а не генерируемых по какому-то алгоритму. Эта программа должна позволить пользователю самостоятельно вводить числа с целью их дальнейшей классификации и размещения числа наблюдений в регистры-счетчики. Очевидно, что первоначально необходимо каким-то образом указать максимальное и минимальное числа вводимого ряда чисел, а затем использовать тот же принцип построения гистограммы, что и в разобранной программе, т. е. изменить «масштаб» этих чисел так, чтобы они оказались в интервале $0 \div 1$.

Следует также обратить внимание еще на один важный аспект работы с программами для ПМК. Для правильного понимания работы программы, т. е. назначения отдельных ее фрагментов, подпрограмм и т. п., необходимо хорошее знание метода, который реализован в используемой программе. Только при условии хорошего знания метода возможна генерация, разработка собственных программ. Однако в том случае, когда пользователь не предполагает тщательного разбора используемых программ, либо их модификации, то знание реализованного в программе метода достаточно на уровне основных идей и разумного понимания и интерпретации получаемых результатов, т. е. на уровне потребителя этого метода. Однако при этом необходимо обращать внимание на те ограничения, которые имеются при использовании тех или иных программ. В противном случае использование программ, особенно чужих, может привести к досадным неприятностям и конфузам.

ГЛАВА 3

ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ ВЫБОРОЧНЫХ МАССИВОВ ДАННЫХ

В обучении примеры важнее правил.

И. Ньютон

3.1. Понятие о генеральной совокупности, выборке, выборочных оценках и их свойствах

Инженер, ученый или технолог, занятые обработкой экспериментальных наблюдений, не всегда имеют возможность использовать для этих целей большое число наблюдений. Это диктуется как сроками выполнения работ, так и рядом других причин. Такое ограничение объемов наблюдений, измерений накладывает, естественно, свою специфику на методы анализа экспериментальных данных и надежность получаемых при этом результатов. Это ограниченное число наблюдений за интересующей нас величиной в математической статистике принято называть выборкой. Например, цехом за смену было изготовлено 1500 доз биологически активного вещества. Для контроля активности изготовленной продукции была взята выборка в количестве 175 доз. Количество наблюдений за интересующей нас величиной, в данном случае активностью препарата, называется объемом выборки.

Тот массив объектов, из которого производится изъятие выборки, называется генеральной совокупностью. В данном случае генеральной совокупностью является 1500 доз препарата, выпущенных цехом за смену. Нередко генеральная совокупность содержит конечное число объектов. Однако если это число достаточно большое, то иногда для упрощения вычислений или теоретических выводов принимают, что генеральная совокупность состоит из бесчисленного множества объектов. Такое допущение при достаточно больших объемах генеральной совокупности практически не сказывается на результатах анализа наблюдений за выборкой. Так, например, изучая выпуск этого же препарата за месяц, квартал или за год, можно генеральную совокупность

считать бесконечной. Объем извлекаемой выборки принято обозначать символом n.

При извлечении элемента выборки из генеральной совокупности, после измерения интересующих нас свойств этого элемента, он может быть возвращен либо не возвращен в генеральную совокупность. По виду отбора выборки различаются на повторные и бесповторные. Повторной называют выборку, при извлечении которой отобранный объект после необходимых измерений над ним возвращается в генеральную совокупность. После чего отбирается следующий объект. Бесповторной называют выборку, в которой отобранный объект не возвращается в генеральную совокупность. На практике обычно используется бесповторный отбор.

Целью извлечения выборки из генеральной совокупности является изучение тех или иных свойств объектов, составляющих выборку, и на основе данных об этих свойствах делается заключение об этих же свойствах, распространяющееся уже на всю генеральную совокупность. Например, после определения активности препарата у выборки из 175 доз можно сделать заключение о средней активности препарата в партии, выпущенной за прошедшую смену. Зная аналогичные данные за предыдущие смены, можно их сравнивать, сопоставлять. Например, можно сравнить средние активности доз, полученные за 1-е смены, за 2-е смены и т. д. Естественно, что достаточно надежно такие данные, называемые выборочными оценками, будут отражать свойства объектов генеральной совокупности лишь в том случае, когда объекты выборки правильно представляют эти свойства. Это требование иначе формулируется следующим образом: выборка должна быть репрезентативной (представительной). В математической статистике доказывается, что выборка будет репрезентативной при условии случайности отбора элементов выборки из генеральной совокупности. Такая случайность достигается в том случае, если все объекты имеют одинаковую вероятность попасть в выборку. Например, если 175 доз были отобраны в качестве выборки из партии объемом 1500 доз только в течение последних 30 минут смены, то полученная выборка не будет репрезентативной.

После получения репрезентативной выборки производится измерение величины какого-то интересующего

нас признака у объектов выборки. Естественно, что полученные при измерении величины могут отличаться друг от друга. При этом чем больше объем выборки, тем более громоздкими и менее наглядными для быстрого восприятия оказываются полученные результаты измерения.

3.1.1. Вариационный ряд и гистограмма

Как уже отмечалось выше, одной из концепций современной прикладной статистики является концепция редукции (свертки) информации. Представление экспериментальных данных в компактной форме значительно облегчает их восприятие, сравнение с ранее полученными результатами и генерацию тех или иных гипотез о модели процесса. В противоположность концепции рандомизации (обеспечения случайности) при получении выборки редукцию полученной информации можно производить только при условии ее систематизации. Первым шагом такой систематизации является упорядочение этих наблюдений и расположение их в порядке возрастания. Расположенные таким образом значения измеренного у объектов выборки признака называют статистическим рядом или вариационным рядом.

Разделим весь имеющийся диапазон величин X на интервалы. Количество интервалов K обычно вычисляют по следующей полуэмпирической формуле Штюргеса:

$$K = 1 + 3{,}32 \cdot \lg n \qquad (3.1.1)$$

с округлением до большего целого, где n — объем выборки. Длина каждого интервала определяется по формуле

$$l = \frac{X_{max} - X_{min}}{K}, \qquad (3.1.2)$$

где X_{max} и X_{min} — соответственно максимальное и минимальное значения признака X. Величину X_{max}—X_{min} называют размахом выборки. Далее определяем середину i-го интервала $\tilde{x}_i$ по формуле

$$\tilde{X}_i = \frac{X_{i+1} + X_i}{2}. \qquad (3.1.3)$$

После этого распределяем имеющиеся в статистическом ряду наблюдения в полученные интервалы. При этом в i-тый интервал относятся данные, удовлетворяющие следующему неравенству:

$$\bar{X}_i - \frac{l}{2} \leqslant X < \bar{X}_i + \frac{l}{2}. \qquad (3.1.4)$$

Таким образом, значения X, лежащие на границе между интервалами, относятся к правому интервалу. Значения X, равные X_{max}, относятся к левому интервалу. Для каждого из интервалов определяем количество элементов выборки (m_i), попавшим в i-тый интервал. Числа m_i называются частотами, а их отношения к объему выборки $\frac{m_i}{n} = p_i$ — относительными частотами попадания в i-тый интервал, или частостями. Очевидно, что

$$\sum_{i=1}^{k} m_i = n, \quad \text{а} \sum_{i=1}^{k} p_i = 1.$$

Проведенные вычисления производят первую редукцию полученной выборочной информации. При этом статистический ряд наблюдений трансформируется в ряд распределения, который можно представить в виде табл. 4.

Таблица 4

Границы интервалов	X_1; X_2	X_2; X_3	...	...	$X_{к-1}$;$X_к$	$X_к$; $X_{к+1}$
Частота m_i	m_1	m_2	...	...	$m_{к-1}$	$m_к$
Частость p_i	p_1	p_2	...	...	$p_{к-1}$	$p_к$

Для большей наглядности и улучшения визуального восприятия статистические ряды распределения удобно представлять в виде гистограмм. Графическое построение гистограммы заключается в следующем. По горизонтальной оси откладываются границы k интервалов. На каждом из интервалов, как на основании, строится прямоугольник. При этом высота h_i каждого такого прямоугольника равна абсолютной (m_i) или относительной (p_i) частоте данного интервала.

Построение гистограмм и их анализ является одним из наиболее простых, доступных и широко распространенных методов анализа экспериментальных данных.

При визуальном восприятии графической гистограммы, особенно если имеется их набор, опытный исследователь может извлечь очень много информации об исследуемом процессе или системе. Для иллюстрации этого обратимся к рис. 1—4. На рис. 1—3 изображены гистограммы распределения толщины гальванического покрытия на деталях за три последовательные недели одного месяца. На рис. 4 дана гистограмма аналогичного распределения за все три недели в целом. Данные, приведенные на этих гистограммах, относятся к деталям, полученным на одной установке, например установке № 1. Сравнительный анализ гистограмм, изображенных на рис. 1—3, показывает, что по своему виду и форме они подобны друг другу. Однако они отличаются тем, что центры распределения, т. е. интервалы с максимальной частотой попадания наблюдений, сдвинуты относительно друг друга. По мере перехода от недели к неделе происходит смещение центра распределения на этих гистограммах вправо, в сторону больших толщин гальванического покрытия. Такие распределения за относительно небольшой период наблюдения называются мгновенными распределениями [12]. Сдвиг, дрейф центра распределения во времени приводит к тому, что на гистограмме суммарного распределения (см. рис. 4) вершина гистограммы стала плоской, т. е. интервал значений толщины слоя, имеющих наибольшую вероятность появления, расширился. Более подробно количественные характеристики эмпирических распределений и их вычисление с помощью ПМК будут рассмотрены в п. 3.2. Плосковершинность суммарного распределения является одним из показателей неустойчивости, нестабильности [12—14] исследуемого процесса во времени. Наличие двух достаточно выраженных вершин на гистограмме может свидетельствовать о том, что исследуемая выборка является суммой, смесью двух отдельных выборок. Например, это может быть сумма двух выборок по установке № 1 и установке № 2 из примера, рассмотренного выше. Причем для каждой из выборок имеется свой устойчивый центр распределения. Либо это может быть выборка по одной установке, например установке № 1, но в какой-то период времени наблюдения произошло резкое смещение центра распределения и в итоге на гистограмме наблюдается две вершины.

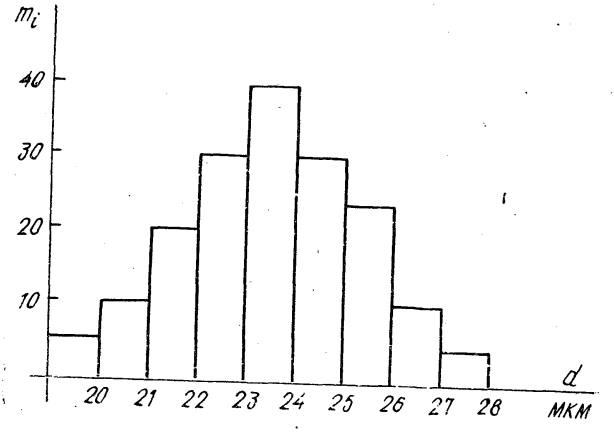

Рис. 1. Гистограмма распределения толщины гальванического покрытия (по данным первой недели)

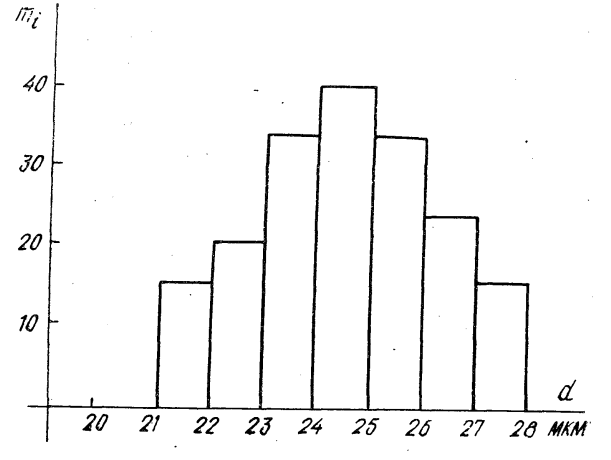

Рис. 2. Гистограмма распределения толщины гальванического покрытия (по данным второй недели)

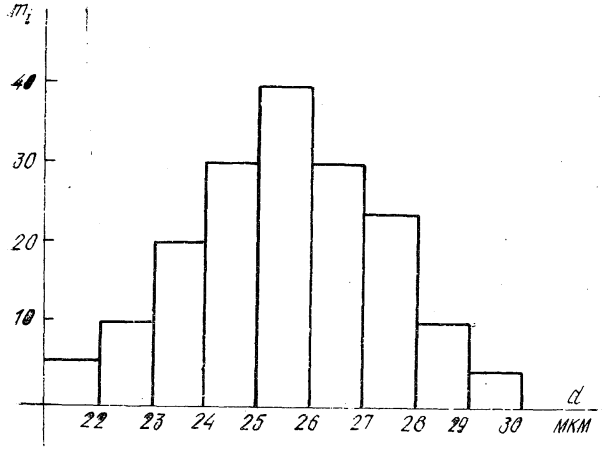

Рис. 3. Гистограмма распределения толщины гальванического покрытия (по данным за третью неделю)

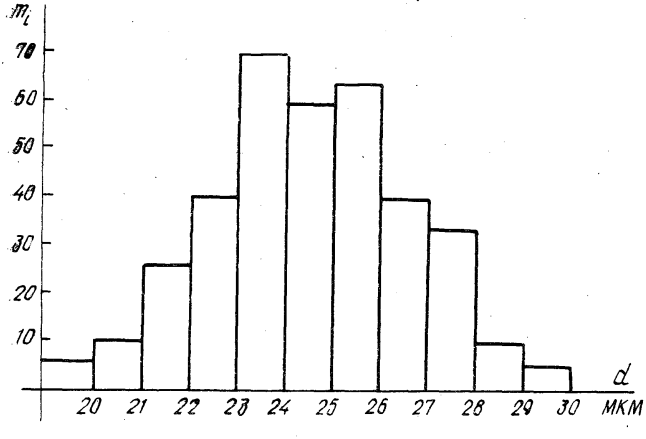

Рис. 4. Гистограмма распределения толщины гальванического покрытия (по данным за три недели в целом)

Гистограммы, изображенные на рис. 5—8, аналогичны предыдущим, только относятся к установке № 2. Сравнение рис. 5—7 говорит об устойчивости, стабильности положения центра распределения. Но отличия

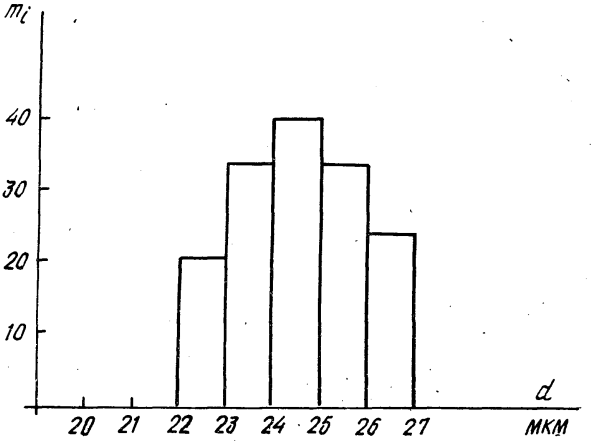

Рис. 5. Гистограмма распределения толщины гальванического покрытия (данные за первую неделю)

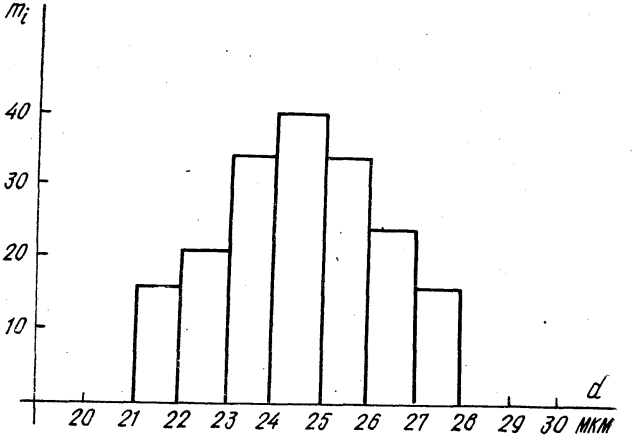

Рис. 6. Гистограмма распределения толщины гальванического покрытия (данные за вторую неделю)

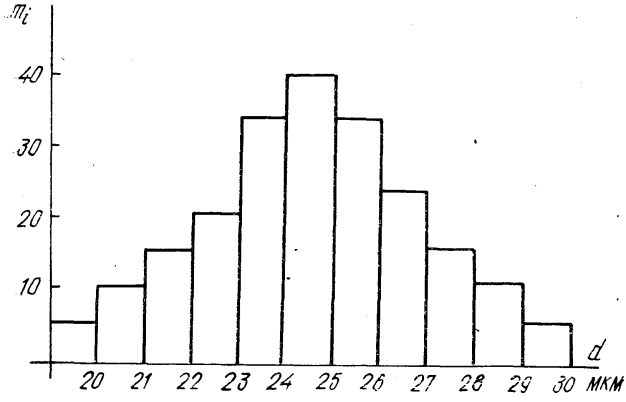

Рис. 7. Гистограмма распределения толщины гальванического покрытия (данные за третью неделю)

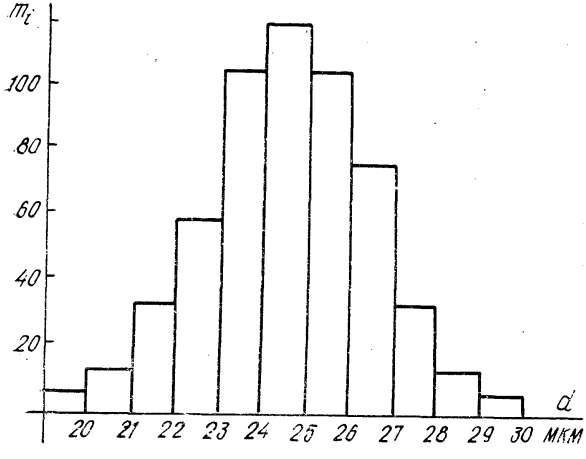

Рис. 8. Гистограмма распределения толщины гальванического покрытия (данные за три недели в целом)

между ними в том, что теперь происходит изменение относительной частоты в центре распределения, т. е. со временем увеличивается разброс толщины. Это приводит к тому, что гистограмма суммарного распределения

(рис. 8) стала островершинной и занимающей больший интервал. Такая островершинность суммарного эмпирического распределения является признаком непостоянства точности [12—14] исследуемого процесса во времени.

Наконец, когда во времени изменяются как положение центра распределения, так и разброс, то гистограмма суммарного распределения может иметь более сложную форму — несимметрическую, с двумя вершинами и т. д. (рис. 9). Конкретный вид гистограммы распределения при этом будет определяться законом изменения во времени как центра распределения, так и разброса. В этом случае можно говорить об отсутствии стабильности процесса и непостоянстве его точности.

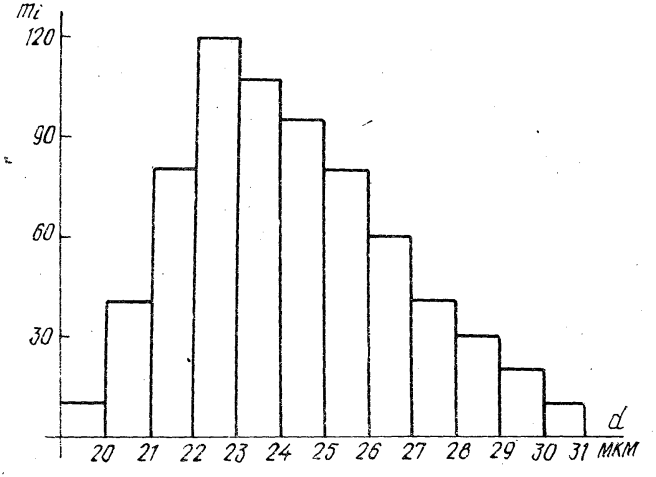

Рис. 9. Гистограмма суммарного распределения, когда во времени изменяется как положение центра распределения, так и разброс

При обработке ряда наблюдений вручную либо с помощью непрограммируемого микрокалькулятора построение гистограммы становится делом трудоемким и утомительным. С помощью ПМК эта процедура сокращается в десятки раз. Ниже приведены несколько программ и инструкций к ним для построения гистограмм.

3.2. Построение гистограмм с помощью ПМК

С помощью ПМК «Электроника Б3-34» можно удобно и быстро решать задачу построения статистического ряда распределения. Приводимые ниже две программы П1/34 и П2/34 позволяют строить ряды распределения с числом интервалов $K \leqslant 10$ либо с $Д = 20$. Кроме нахождения абсолютных значений частоты показания наблюдений в интервалы определяются границы каждого интервала и относительные частоты (частости) для каждого интервала.

При выполнении программы П1/34 в регистрах памяти $1 \div A$ происходит накопление абсолютных частот m_i, в регистре В хранится число задаваемых интервалов $K \leqslant 10$, в регистре С хранится длина одного интервала Δ, а в регистре Д, начиная с 40 адреса, хранится общее число наблюдений, введенных в ПМК.

Пример. При испытании нового рациона кормления цыплят была взята выборка объемом в 40 цыплят. Необходимо построить гистограмму распределения веса цыплят по данной выборке (значения веса цыплят в граммах): 140; 126; 115; 90; 110; 179; 156; 142; 124; 162; 144; 168; 142; 175; 190; 152; 144; 172; 158; 108; 147; 165; 152; 117; 169; 146; 159; 141; 128; 143; 145; 148; 155; 185; 152; 142; 154; 149; 148; 154.

Решение. Согласно инструкции производим ввод программы П2/34, затем набираем число $X_{min} = 90$ и нажимаем клавишу ↑. Набираем число $X_{max} = 190$ и последовательно нажимаем клавиши В/О С/П. После останова индицируется число 1. Вводим первое наблюдение ряда, для этого набираем число 140 и нажимаем клавишу С/П. После останова набираем число 126 и вновь нажимаем клавишу С/П. Аналогичным образом вводим всю выборку. После ввода всей выборки нажимаем клавиши БП 44 С/П. После останова индицируется число 40 — объем введенной выборки, который необходим для контроля числа введенных в ПМК наблюдений ряда и вычисления относительных частот. Выводим результаты обработки по первому слева интервалу, для этого нажимаем клавишу С/П. После останова индицируется относительная частота попадания наблюдений в первый интервал $2{,}5 \cdot 10^{-2}$. Нажимаем клавиши F, и получаем значение абсолютной частоты 1. Вновь нажимаем С/П.

Результат после останова 0. Нажимаем клавиши F: результат 0. Таким образом выводим данные по всем 20 интервалам, которые приведены в табл. 5.

Путем несложных изменений программы П 2/34 можно получать значения относительных частот P_i в процентах (попытайтесь сделать это самостоятельно).

Таблица 5

Номер интервала i	1	2	3	4	5
$P_i = \frac{m_i}{n}$	$2{,}5 \cdot 10^{-2}$	0	0	$2{,}5 \cdot 10^{-2}$	$2{,}5 \cdot 10^{-2}$
m_i	1	0	0	1	1
Номер интервала i	6	7	8	9	10
$p_i = \frac{m_i}{n}$	$5 \cdot 10^{-2}$	$2{,}5 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	0	0
m_i	2	1	2	0	0
Номер интервала i	11	12	13	14	15
$p_i = \frac{m_i}{n}$	$2{,}5 \cdot 10^{-1}$	$1{,}5 \cdot 10^{-1}$	$1{,}25 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$2{,}5 \cdot 10^{-2}$
m_i	8	6	5	4	1
Номер интервала i	16	17	18	19	20
$p_i = \frac{m_i}{n}$	$7{,}5 \cdot 10^{-2}$	$2{,}5 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	0	$5 \cdot 10^{-2}$
m_i	3	1	2	0	2

3.3. Оценка основных параметров выборки (выборочных характеристик)

Как уже говорилось выше, гистограммы, графически изображающие одномерное распределение частот, отличаются друг от друга формой и местом расположения на оси переменной. Для характеристики этих особенностей одновременного распределения частот используются следующие меры:

1. **Меры положения** — меры среднего положения распределения: среднее арифметическое, медиана, мода.

2. **Меры рассеяния** — меры, которые характеризуют изменчивость распределения: дисперсия, стандартное отклонение, размах, коэффициент вариации.

3. **Меры формы** — меры, которые характеризуют отклонение распределения от нормального закона: коэффициенты асимметрии и эксцесса.

Перечисленные выше меры рассматривают как характеристики, существующие в генеральной совокупности. В этом случае такие характеристики принято называть параметрами распределения вероятностей генеральной совокупности. Выборочные же, эмпирические значения характеристик называют оценками или статистиками. Относительные частоты (частости) по своему смыслу также являются выборочными оценками вероятности в генеральной совокупности.

Параметры генеральной совокупности в литературе по статистике принято обозначать буквами греческого алфавита, а их выборочные оценки (статистики) — соответствующими буквами латинского алфавита.

Рассмотрим методы получения выборочных оценок перечисленных выше мер, программы их расчета на ПМК и возможности их применения в исследовательской практике. Отметим, что в дальнейшем мы будем рассматривать только непрерывно распределенные переменные, т. е. непрерывные значения переменных, в противоположность дискретным значениям.

При наличии выборки наблюдений за переменной X объемом n среднее арифметическое значение данной переменной для данной выборки находится по формуле

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i;\ i = \overline{1,\ n}. \quad (3.3.1)$$

Следовательно, величина X является выборочным средним, оценкой среднего значения этой переменной в генеральной совокупности (генерального среднего). Генеральное среднее обычно принято обозначать через μ. Для среднего значения употребляются также другие обозначения: $M\{X\}$, $E\{X\}$.

При симметричном распределении среднее арифметическое $\bar{X}$ находится примерно в центре диапазона из-

менений переменной. Если же распределение асимметричное, т. е. с одной стороны гистограммы наблюдается резкое изменение частот, а с другой — медленное, то значение $\bar{X}$ сдвинуто в одну из сторон от центра диапазона. При этом большая часть наблюдений расположена с одной стороны от среднего, тогда как меньшая часть наблюдений расположена на большом удалении с другой стороны. В этом случае полезно использовать медиану $\bar{X}$. Медиана $\bar{X}$ равна значению переменной, которая делит распределение на 2 равные части так, что каждая содержит по 50% всех наблюдений. Медиана соответствует тому члену упорядоченного вариационного ряда, который делит ряд пополам. В том случае, когда среднее арифметическое $\bar{X}$ располагается на числовой оси справа от медианы $\bar{X}$, т. е. $\bar{X} > X$ и $\bar{X} - X > 0$, принято говорить о положительной асимметрии распределения. Положительно асимметричное распределение имеет отчетливую вытянутость вправо. Наиболее часто встречающееся выборочное значение переменной называют модой. При одновершинном распределении, или унимодальном, моде отвечает одно значение переменной. При многовершинном распределении имеется несколько мод. Такие распределения называют бимодальными (2 моды и вершины), тримодальными (3 моды и вершины). В общем случае принято говорить о мультимодальном распределении.

3.3.1. Понятие о законе больших чисел и центральной предельной теореме

Прежде чем перейти к мерам рассеяния, рассмотрим сущность двух теорем теории вероятностей.

В 1866 году великий русский ученый П. Л. Чебышев — создатель русской школы теории вероятности, опубликовал работу «О средних величинах», в которой доказал одну из этих теорем, названную впоследствии его именем. Эта теорема, называемая также законом больших чисел, имеет огромное значение для практики.

Когда мы говорим случайная величина, то подразумеваем при этом только то, что вследствие многих случайных причин мы не можем априори, т. е. до измерения этой величины, предсказать точное ее количественное значение. В самом деле, например, в технологическом

процессе штамповки деталей на автомате можно перечислить множество таких случайных причин, которые влияют на качество детали (габариты, вес, качество поверхности детали): колебания напряжения в электросети, вследствие чего наблюдаются колебания усилия в прессе, что влечет за собой изменение габаритов детали и т. п.; изменение свойств металла, из которого штампуются детали; изменения режимов настройки автомата от смены к смене, степень износа автомата и т. п. Вследствие этих многочисленных причин только после окончания процесса штамповки детали можно измерить ее размеры и другие показатели качества. Но и на этом этапе имеется не меньшее количество случайных причин, также влияющих на значение измеряемой величины. Поскольку число случайных причин в каждом конкретном случае весьма велико и заранее нельзя установить и предсказать, в какой комбинации и последовательности эти причины повлияют на значение исследуемой переменной, то казалось бы невозможно установить закономерности поведения достаточно большого числа случайных величин. Однако это не так. При достаточно общих условиях суммарное поведение большого числа случайных величин практически утрачивает случайный характер и проявляется закономерность. Естественно, что для практической деятельности важно знать тот комплекс условий, при выполнении которых совместное действие этого множества случайных причин практически не повлияет на конечный результат. Эти условия и рассматриваются в теоремах, которые носят название закона больших чисел. Теорема, которую П. Л. Чебышев опубликовал в 1866 г., и является наиболее общей формой закона больших чисел. Существо этой теоремы можно сформулировать так: несмотря на то, что отдельные независимые друг от друга случайные величины могут иметь значения, далекие от генерального среднего, среднее арифметическое (выборочная оценка генерального среднего) достаточно большого числа случайных величин с большой вероятностью принимает значения, близкие к генеральному среднему. Иначе говоря, хотя отдельные случайные величины могут иметь значительный разброс, среднее арифметическое имеет гораздо меньший разброс, т. е. оно утрачивает характер случайной величины. Широко используемый в практике выбо-

рочный метод и использует это свойство для оценки параметра генеральной совокупности. Однако каковы же условия, при которых эта особенность реализуется на практике? На этот вопрос также отвечает теорема Чебышева. Рассмотрим результаты измерения веса штампованных деталей: P_1, P_2, ..., P_n. Теорему Чебышева можно применить к этим случайным величинам при выполнении следующих трех условий: а) эти величины между собой попарно независимы; б) принадлежат одной и той же генеральной совокупности и имеют одно генеральное среднее; в) отклонения этих величин от генерального среднего равномерно ограничены. Первое условие выполняется, когда результат каждого отдельного измерения (наблюдения) не зависит от всех остальных результатов. Второе условие означает также и отсутствие систематической погрешности у измерительного прибора. Третье условие означает ограниченность погрешности измерения. При выполнении этих условий среднее арифметическое достаточно большого числа измерений весьма мало отличается от генерального среднего значения веса штампованной детали.

Вторая теорема Чебышева (так называемая центральная предельная теорема) была опубликована им в 1887 г. в «Записках Академии наук». В 1900 г. выдающийся русский математик А. М. Ляпунов доказал центральную предельную теорему в более общей, чем у П. Л. Чебышева, формулировке. Следствие из центральной предельной теоремы гласит: если случайная величина X представляет из себя сумму очень большого числа взаимно независимых случайных величин, каждая из которых ничтожно влияет на всю сумму, то X имеет распределение, близкое к нормальному. При этом законы распределения величин, составляющих сумму, могут быть неизвестны. Один из параметров, используемый при аналитической записи выражения для нормального закона распределения, уже рассмотрен: это генеральное среднее и его выборочная оценка — среднее арифметическое. Вторым параметром является дисперсия. Этот термин происходит от латинского dispengo — рассыпать, рассеивать, разбрасывать. Понятие дисперсии ввел в 1877 г. немецкий ученый Лексис. Дисперсия является мерой рассеяния отдельных значений случайной величины совокупности вокруг среднего значения этой совокупно-

сти. Дисперсия представляет собой среднее значение квадрата отклонений случайной величины от среднего значения совокупности. Выборочная оценка дисперсии вычисляется по формуле

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. \qquad (3.3.2)$$

Положительное значение корня квадратного из дисперсии называется средним квадратичным или стандартным отклонением (ошибкой), а иногда просто стандартом и обозначается символом s:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}. \qquad (3.3.3)$$

Относительная квадратическая ошибка, выраженная в процентах от среднего значения случайной величины, называется коэффициентом вариации и обозначается символом v:

$$v = \frac{s}{\bar{X}} \cdot 100. \qquad (3.3.4)$$

Интервал изменения значения случайной величины в выборке от минимального до максимального называется размахом и обозначается символом R:

$$R = X_{max} - X_{min}. \qquad (3.3.5)$$

3.3.2. Статистики нормального закона распределения

При выполнении условий центральной предельной теоремы значения случайной величины распределяются по нормальному закону, плотность распределения вероятностей которого определяется формулой

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad (3.3.6)$$

где μ — генеральное среднее; σ^2 — генеральная дисперсия.

Отклонение от нормального закона проявляется в изменении формы кривой плотности распределения вероятностей и формы гистограммы. Для описания формы в статистике используют две формы: коэффициент асим-

метрии и коэффициент эксцесса. Коэффициент асимметрии A^* вычисляют по формуле

$$A^* = m_3 \cdot m_2^{-3/2}, \qquad (3.3.7)$$

а коэффициент эксцесса E^* по формуле

$$E^* = m_4 \cdot m_2^{-2} - 3. \qquad (3.3.8)$$

В этих двух формулах через m_k обозначены так называемые центральные моменты k-го порядка, вычисляемые по следующей формуле:

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^k. \qquad (3.3.9)$$

Для нормального распределения теоретические значения обоих коэффициентов A^* и E^* равны нулю. Естественно, что выборочные оценки этих коэффициентов могут и отличаться от нуля, даже если выборка взята из генеральной совокупности, в которой исследуемая величина подчиняется нормальному закону распределения вероятностей. Статистики A^* и E^*, вычисляемые по формулам (3.2.7) и (3.2.8), являются смещенными выборочными оценками соответствующих генеральных параметров распределения. Несмещенную оценку генеральных параметров получают путем умножения смещенной оценки на некоторую функцию от объема выборки n.

Для коэффициентов асимметрии и эксцесса такие несмещенные оценки определяют по формулам:

$$A = \frac{\sqrt{n(n-1)}}{n-2} \cdot A^* \qquad (3.3.10)$$

и

$$E = \frac{n-1}{(n-2)\cdot(n-3)} [(n+1)\cdot E^* + 6]. \qquad (3.3.11)$$

Все рассмотренные выше параметры выборок представляют собой оценки соответствующих параметров генеральной совокупности, т. е. генеральных параметров. Поскольку каждая из этих оценок получена в виде числа, т. е. одной точки на числовой оси, то такие оценки принято называть точечными. Очевидно, что в силу множества случайных причин в выборку, которой мы опе-

---

* Звездочка означает смещенность статистики.

рируем при вычислении этих точечных оценок параметров (статистик), случайно попали именно те, а не иные элементы генеральной совокупности. Поэтому можно говорить о случайном характере полученного значения статистики. Если из генеральной совокупности взять повторную выборку того же самого объема, что и у предыдущей выборки, и вновь вычислить точечные оценки тех же самых параметров, например, среднее арифметическое, стандарт, коэффициенты асимметрии и эксцесса, то мы получим уже другие значения этих оценок. Та же самая картина получится и для 3-й, 4-й и т. д. выборок. Вычисляя для каждой из выборок $\bar{X}$, s, A и E, мы получим следующие четыре ряда распределений:

1-й ряд: $\bar{X}_1$; $\bar{X}_2$; $\bar{X}_3$, ..., $\bar{X}_n$;
2-й ряд: s_1; s_2; s_3; ..., s_n;
3-й ряд: A_1; A_2; A_3; ..., A_n;
4-й ряд: E_1; E_2; E_3; ..., E_n.

Если для каждого из этих рядов найти свои аналогичные параметры, то можно получить значительно более точную и надежную информацию о параметрах генеральной совокупности. При этом дисперсия распределения средних значений выборок будет связана с дисперсией самой величины в генеральной совокупности следующим образом:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}, \qquad (3.3.12)$$

а среднеквадратические отклонения (стандарты) как

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}. \qquad (3.3.13)$$

Здесь $s_{\bar{x}}$ — стандартное отклонение средних значений выборок; s — стандартное отклонение значений случайной величины; n — объем выборки. Из формулы (3.3.13) следует, что стандартное отклонение средних значений выборок в $1/\sqrt{n}$ меньше, чем стандартное отклонение случайной величины в генеральной совокупности. Это свойство можно использовать для повышения точности оценки среднего значения. Например, для увеличения точности оценки генерального среднего вдвое объем выборки необходимо увеличивать в 4 раза, втрое — в 9 раз, на порядок — в 100 раз.

Для нормально распределенной генеральной совокупности стандартные отклонения для второго и третьего рядов, т. е. для коэффициентов асимметрии и эксцесса, зависят только от объема выборки и определяются следующими выражениями:

$$s_A = \sqrt{\frac{6(n-1)}{(n+1)\cdot(n+3)}}, \qquad (3.3.14)$$

$$s_E = \sqrt{\frac{24n\cdot(n-2)\cdot(n-3)}{(n+1)^2(n+3)\cdot(n+5)}}. \qquad (3.3.15)$$

Программа П3/34 предназначена для обработки выборки наблюдений и вычисления среднего арифметического $\bar{X}$, стандартного отклонения s, коэффициента вариации v и ошибки среднего (стандартного отклонения среднего) $s_{\bar{x}}$. Для контрольной проверки программы П3/34 удобно использовать выборку 1; 2; 3. По окончании работы П3/34 имеем $\bar{X}=2$; $s=1$; $s_{\bar{x}}=5{,}7735027\cdot10^{-1}$, $v=50\%$ (коэффициент вариации программой П3/34 вычисляется в процентах).

В том случае, когда имеется необходимость вычислять не только конечные оценки $\bar{X}$, s, v и $s_{\bar{x}}$ для всей имеющейся выборки наблюдений, но и промежуточные оценки для какой-то введенной части наблюдений, также можно использовать программу П3/34. Такая необходимость может возникнуть, например, в том случае, когда $s_{\bar{x}}$ достигает определенной величины, сравнимой с погрешностью средств измерений, либо представляет интерес проследить поведение оценок в зависимости от объема выборки. В этом случае после окончания работы программы П3/34 на индикатор вызываются значения $\bar{X}$, s, v и $s_{\bar{x}}$ последовательным нажатием ИПА, ИПВ, ИПС, ИПД (результаты записываются), затем нажимаются клавиши БП 06, вводится очередное значение переменной из выборки и нажимается С/П (по п. 2 и п. 3 инструкции), далее все вновь по инструкции. Например, после ввода выборки 1, 2, 3 и получения соответствующих оценок необходимо ввести следующие элементы выборки: 4, 5. Нажимаем БП 06 4 С/П. После останова набираем 5 и нажимаем С/П. После очередного останова нажимаем клавиши $\overrightarrow{\text{ШГ}}$ и С/П. После останова извлекаем

оценки из регистров А, В, С и Д: ИПА : $\bar{X}=3$; ИПВ; $s=1{,}5811388$; ИПС : $v=52{,}704626\%$; ИПД : $s_{\bar{x}}=$ $=7{,}0710679\cdot10^{-1}$.

В исследовательской практике очень часто в течение некоторого периода производят обработку отдельных небольших выборок. Например, за две недели были обработаны 20 выборок, суммарный объем которых равен 600 наблюдениям. Помимо оценок среднего, дисперсии и других параметров по отдельным выборкам нас интересуют эти же параметры суммарной выборки из 600 наблюдений. Программа П4/34 позволяет определить оценки этой суммарной выборки, не вводя повторно все 600 наблюдений вместе. В этом случае необходимо по каждой из отдельных выборок знание следующих параметров: объема выборки n_i, оценки среднего арифметического $\bar{X}_i$ и стандартного отклонения s_i.

Пример. В табл. 6 представлены данные о размерах эритроцитов крови, полученные на четырех мазках от одной мыши [43, с. 116]. Требуется определить уточненные оценки параметров по суммарной выборке объемом в $392+364+345+448=1549$ эритроцитов.

Таблица 6

Параметры	1-й мазок	2-й мазок	3-й мазок	4-й мазок
Объем выборки n_i	392	364	345	448
Средний диаметр $\bar{X}_i$, мкм	8,33	8,26	8,30	8,28
Дисперсия s^2_i, мкм^2	0,850	0,681	0,706	0,861

Решение. После ввода программы П4/34 набираем число 392 и нажимаем клавишу ↑. Затем набираем число 8,33 и вновь нажимаем клавишу ↑. Далее набираем число 0,850, но не нажимаем клавиши В/О С/П согласно инструкции к П4/34, т. к. число 0,850 есть дисперсия s^2, а не стандартное отклонение s. Поэтому предварительно извлекаем квадратный корень, нажав клавиши $F\sqrt{\ }$, и только затем нажимаем клавиши В/О С/П. После останова на ПМК индицируется число 27532,799. Аналогично вводим данные 2-й выборки, В/О С/П — результат 52614,848, 3-й — результат 76624,762 и 4-й — результат 79678,355. Далее нажимаем С/П — результат $\bar{\bar{x}}=$

$=8{,}296786$, С/П — результат $s=8{,}672267\cdot10^{-1}$, С/П — результат $v=10{,}452561\%$, С/П — результат $s_{\bar{x}}=2{,}5640095\cdot10^{-2}$.

В том случае, когда необходимо получить оценки $\bar{X}$, s, A и E по выборке, по которой ранее уже была построена гистограмма, то для этого удобно воспользоваться программой П5/34 [7,62]. Данная программа обрабатывает выборку, в которой наблюдение сгруппировано по интервалам равной длины h. При этом в качестве новых данных вводятся в ПМК значения центров интервалов X_i, соответствующих абсолютных частот попадания наблюдений в эти интервалы m_i и длина интервалов h. Следует учитывать тот факт, что формулы, по которым производятся вычисления, обеспечивают приближенное равенство. Точность оценок повышается по мере роста объема выборки и уменьшения длины интервала (увеличения количества интервалов). Программа П5/34 обеспечивает вычисление начальных и центральных моментов с учетом поправок Шеппарда.

После вычисления центральных моментов по программе П5/34 несложно вычислить оценку коэффициента асимметрии $A^*=m_3/\sqrt{m_2^3}$ и $E^*=m_4/m_2^2-3$, напомним, что $\bar{X}=m_1$. $s^2_*=m_2$.

Для вычисления A^* после окончания работы П5/34 необходимо нажать клавиши: ИП3 ИП4↑Fx²×F√÷ПА. Аналогично для вычисления E^* : ИП2 ИП4 Fx²÷3—ПВ. Отметим также, что П5/34 дает смещенную оценку s_*^2, для получения несмещенной оценки $s^2=\frac{s_*^2\cdot n}{n-1}$, нажимаем клавиши в следующей последовательности: ИП6 ИП4 ИП4 1 — ÷×. Если требуется оценка стандартного отклонения, дополнительно нажимаются клавиши F√.

Для обработки несгруппированных данных ниже приводится программа П6/34, которая позволяет также определить значения $\bar{X}$, s, A^* и E^* [7]. Программа П6/34 дает несмещенную оценку дисперсии s, которая хранится по окончании работы П6/34 в седьмом регистре.

3.4. Оценка доверительных и толерантных интервалов

3.4.1. Интервальные оценки

Как уже говорилось выше, выборочные оценки генеральных параметров, которые определяются одним числом, называются точечными оценками. Значения этих оценок, вычисленные на основе полученной выборки, в силу случайности выборки являются случайными. При многократном отборе выборок из генеральной совокупности и вычислении выборочных оценок параметров по ним можно заметить, что эти оценки, например оценки генерального среднего, располагаются на числовой оси вблизи друг друга и занимают некоторый интервал на числовой оси. Помимо точечных оценок в практике широко используется метод интервальных оценок параметров генеральной совокупности. Интервальной называют такую оценку, которая определяется двумя числами — концами интервала. Интервальная оценка по своей сути является более информативной, поскольку она содержит в себе также точность и надежность этой оценки.

Например, по данным одной выборки была получена выборочная оценка C^* генерального параметра C генеральной совокупности. Чем точнее получена выборочная оценка C^*, тем меньше модуль разности $|C-C^*|$. Обозначим модуль разности оценки и истинного значения генерального параметра через δ. Это число δ и характеризует точность оценки генерального параметра. Если $\delta>0$ и $|C-C^*|<\delta$, то меньшим δ отвечают и более точные оценки параметра. Это неравенство $|C-C^*|<\delta$ не является абсолютным. В математической статистике можно лишь говорить о вероятности β, с которой это неравенство действительно осуществляется.

Эту вероятность β принято называть доверительной вероятностью. Как правило, значение β задается самим исследователем, причем значение β берут близким к 1. Наиболее часто в практике пользуются значениями β, равными 0,95; 0,99; 0,999. Например, выберем для оценки среднего арифметического $\bar{X}$ значение доверительной вероятности $\beta=95\%$, т. е. 0,95. Тогда это означает, что вероятность выполнения неравенства $|\mu-\bar{X}|<\delta$ равна 95%. Обозначив вероятность через P, то же самое запишем иначе

$$P[|\mu-\bar{X}|<\delta]=0{,}95. \qquad (3.4.1)$$

Заменим неравенство $|\mu-\bar{X}|<\delta$ идентичным ему выражением $-\delta<\mu-\bar{X}<\delta$, или $\bar{X}-\delta<\mu<\bar{X}+\delta$, получим следующее выражение:

$$P[\bar{X}-\delta<\mu<\bar{X}+\delta]=0{,}95. \qquad (3.4.2)$$

Это выражение следует понимать следующим образом: вероятность того, что величина неизвестного генерального среднего μ заключена внутри интервала $(\bar{X}-\delta, \bar{X}+\delta)$, равна 95%. При этом доверительным называется интервал $(C^*-\delta, C^*+\delta)$, который заключает внутри себя (покрывает) неизвестное значение генерального параметра с заданной доверительной вероятностью (надежностью) β. Границы интервала называют в этом случае доверительными границами. Доверительную вероятность называют также иногда уровнем достоверности. Величину $\alpha=1-\beta$ называют уровнем значимости. Для рассмотренного выше примера уровень значимости $\alpha=5\%=0{,}05$.

Метод оценок генеральных параметров с помощью доверительных интервалов был разработан американским статистиком Дж. Нейманом и английскими статистиками Р. Фишером и Э. С. Пирсоном [15]. Ниже мы рассмотрим методы нахождения доверительных интервалов для следующих статистик: среднего, дисперсии, коэффициента вариации и медианы для случая нормального распределения.

3.4.2. Доверительные интервалы для среднего

Рассмотрим процедуру нахождения двухстороннего доверительного интервала для оценки генерального среднего совокупности, имеющей нормальное распределение количественного признака с известным значением дисперсии σ^2. В этом случае искомый доверительный интервал при заданном значении σ и доверительной вероятности β определяются по формуле

$$\bar{X}-z_{\frac{1+\beta}{2}}\cdot\frac{\sigma}{\sqrt{n}}<\mu<\bar{X}+z_{\frac{1+\beta}{2}}\cdot\frac{\sigma}{\sqrt{n}}, \qquad (3.4.3)$$

где $\bar{X}$ — выборочное среднее; n—объем выборки; $z_{\frac{1+\beta}{2}}$ — квантиль нормального распределения [16—19]. Очень

важно помнить, что в данном случае речь идет о двухстороннем интервале, ограниченном слева и справа. Поэтому используется квантиль $z_{\frac{1+\beta}{2}}$, а не квантиль z_β.

Поясним это несколько подробнее. Например, мы задались значением доверительной вероятности $\beta=90\%$, т. е. хотели бы найти такие границы этого интервала, внутри которых расположилось бы 90% всех значений выборочных средних при многократном извлечении выборок из генеральной совокупности. При этом нас могут интересовать три различных случая ограничения интервала: 1) ограничение слева, т. е. ограничение минимальным значением, при этом граница справа $+\infty$; 2) ограничение справа, т. е. ограничение максимальным значением, при этом граница слева — ∞; 3) ограничение как слева, так и справа. Первые два случая отвечают односторонним (левосторонний и правосторонний) доверительным интервалам. В этих случаях доверительные интервалы будут выглядеть соответственно следующим образом:

$$\bar{X} - z_{0,9} \cdot \frac{\sigma}{\sqrt{n}} < \mu < +\infty, \qquad (3.4.4)$$

$$-\infty < \mu < \bar{X} + z_{0,9} \cdot \frac{\sigma}{\sqrt{n}}. \qquad (3.4.5)$$

Для таких интервалов хвост распределения вероятностей ограничивается значением β слева или справа, т. е. возможные ошибки (выход выборочных средних за установленную лево-, правостороннюю границу) будут иметь вероятность $1-\beta=\alpha$. В практике же гораздо чаще используются двухсторонние доверительные интервалы. В этом случае значение возможной ошибки α (уровень значимости) для каждой из 2 границ делится пополам. И тогда вместо квантиля $z_{0,9}$ используется квантиль $z_{\frac{1+0,9}{2}} = z_{0,95}$. Теперь вероятность выхода выборочных средних за каждую (левую или правую) из границ составит уже не 10% как раньше, а только 5%. В различной литературе по статистике встречаются как выражения $z_{\frac{1+\beta}{2}}$, так и $z_{1-\frac{\alpha}{2}}$. Во избежание путаницы рассмотрим связь между этими обозначениями.

Как уже говорилось выше, $\alpha = 1 - \beta$ — уровень значимости, т. е. это тот уровень, при превышении которого мы считаем происходящее событие значимым, не случайным. И напротив, если вероятность происшедшего события меньше уровня значимости, то такое событие мы считаем незначимым, случайным. Легко видеть, что

$$\frac{1+\beta}{2} = \frac{1+(1-\alpha)}{2} = \frac{2-\alpha}{2} = 1 - \frac{\alpha}{2}.$$

Это обозначает, что две записи $z_{\frac{1+\beta}{2}}$ и $z_{1-\frac{\alpha}{2}}$ по своему смыслу идентичны; при

$$\beta = 0{,}9 \quad \alpha = 0{,}1 \text{ и } z_{\frac{1+0{,}9}{2}} = z_{1-\frac{0{,}1}{2}} = z_{0{,}95}.$$

В табл. 7 приведены значения квантилей $z_{\frac{1+\beta}{2}}$ для наиболее часто употребляемых значений β [18].

Таблица 7

β	$z_{\frac{1+\beta}{2}}$	β	$z_{\frac{1+\beta}{2}}$	β	$z_{\frac{1+\beta}{2}}$
0,80	1,282	0,86	1,475	0,99	2,576
0,81	1,310	0,87	1,513	0,999	3,291
0,82	1,340	0,88	1,554	0,6826	1,0
0,83	1,371	0,89	1,597	0,9544	2,0
0,84	1,404	0,90	1,645	0,9973	3,0
0,85	1,439	0,95	1,960	0,999936	4,0

Из выражения (3.4.3) вытекает несколько следствий, на которых необходимо отдельно остановиться. Во-первых, поскольку полуширина доверительного интервала $\delta = z \cdot \sigma/\sqrt{n}$, то, следовательно, при возрастании объема выборки n δ уменьшается, а точность оценки возрастает. Во-вторых, увеличение доверительной вероятности приводит к увеличению $z_{\frac{1+\beta}{2}}$ и увеличению δ, что означает уменьшение точности оценки. Это противоречие кратко формулируют следующим образом: надежное высказывание некатегорично; категоричное высказывание ненадежно. В самом деле, если мы зададимся дове-

рительной вероятностью 100% (категоричное высказывание), то в этом случае доверительный интервал будет лежать в границах $-\infty$, $+\infty$ (ненадежность оценки по категоричному высказыванию).

Для иллюстрации применения выражения (3.4.3) рассмотрим следующий пример. На 36 делянках опытного хозяйства с примерно одинаковыми почвами выращивается районированный сорт пшеницы. Известно, что распределение урожайности подчиняется нормальному закону в данном районе почв и климата с известным стандартным отклонением $\sigma=1{,}8$ ц/га. Найти доверительный интервал для генерального среднего с доверительной вероятностью $P=95\%$, если выборочное среднее равно 28 ц/га. Согласно (3.4.1) определяем точность оценки среднего $\delta=z\cdot\sigma/\sqrt{n}$: $1{,}96\cdot1{,}8/\sqrt{36}=0{,}588$. Таким образом, нижняя граница доверительного интервала $\bar{d}-\delta=28-0{,}588=27{,}412\approx27{,}4$ ц/га, верхняя — $28+0{,}588=28{,}588\approx28{,}6$ ц/га. Полученные оценки доверительного интервала, а точнее его границы, являются теми границами, внутри которых будут находиться 95% значений выборочного среднего урожая, если произвести достаточно большое число выборок. При этом лишь в 5% выборочное значение среднего урожая выйдет за границы доверительного интервала.

В практических исследованиях гораздо чаще точное значение дисперсии распределения неизвестно. В этом случае при построении доверительного интервала для среднего используют уже не одну выборочную точечную оценку — среднее арифметическое, но и другую выборочную точечную оценку — выборочный стандарт. Общее выражение для нахождения границ доверительного интервала аналогично выражению (3.4.3), за тем лишь исключением, что вместо квантиля нормального распределения z используется квантиль так называемого t-распределения Стьюдента. Английский химик В. С. Госсет, работавший на пивоваренных заводах Гиннеса, занялся проблемой оценки генеральных параметров на основе малых выборок ($n\leqslant30$). К этому его побудила необходимость оценки результатов химического анализа по малому числу проб. В 1908 г. В. С. Госсет публикует свои результаты по статистике малых выборок в журнале «Биометрика» под псевдонимом «Студент» (Student). В 1924 г. t-распределение Стьюдента было уточнено из-

вестнейшим английским статистиком Р. А. Фишером на международном математическом конгрессе в г. Торонто. Это распределение произвело революцию в теории малых выборок. Кривая плотности t-распределения напоминает кривую нормального распределения, но зависит от объема выборки n. Число $f=n-1$ называется числом степеней свободы t-распределения. При числе степеней свободы $25 \div 30$ и доверительной вероятности $P_{дов}=90 \div 95\%$ значения квантилей t и z приблизительно совпадают, а при $f \to \infty$ t-распределение асимптотически переходит в нормальное распределение. Таблицы t-распределения, которыми пользуются при расчетах, приведены в [17, 18].

Заменив в (3.4.3) z на $t_{n-1;\, 1-\frac{\alpha}{2}}$, а σ на s, получим

$$\delta = \pm t_{n-1;\, 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}. \qquad (3.4.6)$$

При числе степеней $f \gg 20 \div 30$ формулу (3.4.6) можно заменить выражениями:

$$\delta = \pm 2s \text{ и } \delta = \pm 3s \qquad (3.4.7)$$

соответственно для доверительной вероятности, равной 95 и $99 \div 99{,}7\%$. В таких случаях эти доверительные интервалы называют двухсигмовым и трехсигмовым соответственно. Распределение Стьюдента симметрично, поэтому и границы доверительного интервала также оказываются симметричными относительно выборочного среднего. Относительно искомого генерального среднего можно утверждать, что в среднем у $100 \cdot \beta\%$ всех выборок между найденными границами заключено истинное значение генерального среднего, и наоборот, у $100 \cdot \alpha\%$ всех выборок неизвестный параметр не накрывается этим интервалом. В среднем $100 \cdot (1-\beta)\%$ всех выборок имеют параметр вне этих границ, причем у $100 \cdot \frac{\alpha}{2}\%$ имеют параметр слева и у такого же количества выборок — справа, т. к. $\frac{\alpha}{2} + \beta + \frac{\alpha}{2} = 1$. Для иллюстрации этого метода рассмотрим следующий пример. На детали произведено четыре измерения толщины

осажденного защитного покрытия: 7,8; 8,2; 8,0; 8,1 мкм. Определить доверительный интервал для средней толщины защитного покрытия, используя: а) 2 измерения; б) 3 измерения; в) 4 измерения при $\beta=90\%$ ($\alpha=10\%$):

а) $\bar{d}_1=8$ мкм; $n_1=2$; $f_1=1$; $t_{1;\,0,05}=6{,}31$; $s_1=0{,}28$ мкм;

$\delta_1=1{,}25$ мкм; $\bar{d}=\bar{d}_1\pm1{,}25$ $(6{,}75\div9{,}25)$ мкм.

б) $\bar{d}_2=8$ мкм; $n_2=3$; $f_2=2$; $t_{2;\,0,05}=2{,}92$; $s_2=0{,}2$ мкм;

$\delta_2=0{,}34$ мкм; $\bar{d}=\bar{d}_2\pm0{,}34$ $(7{,}66\div8{,}34)$ мкм.

в) $\bar{d}_3=8{,}025$ мкм; $n_3=4$; $f_3=3$; $t_{3;\,0,05}=2{,}35$; $s_3=0{,}17$ мкм;

$\delta_3=0{,}2$ мкм; $\bar{d}=\bar{d}_3\pm0{,}2$ $(7{,}8\div8{,}2)$ мкм

Как видно из этих расчетов, величина доверительного интервала наиболее резко падает при переходе от двух к трем параллельным измерениям. Отсюда следует вывод, что разумно проводить не менее трех параллельных измерений, что обеспечивает число степеней свободы $f=n-1>1$.

В табл. 8 приведены значения $t_{f;\,\alpha}$ для $f=1$; 2; 3 для иллюстрации того, что наиболее резко величина квантиля t-распределения Стьюдента изменяется при переходе от $f=1$ к $f=2$.

3.4.3. Определение числа наблюдений и доверительной вероятности

Из выражений (3.4.3) и (3.4.6) вытекает 2 важных для практики следствия. Рассмотрим их подробнее. Из равенства

$$\delta = z\cdot\frac{\sigma}{\sqrt{n}} \qquad (3.4.8)$$

следует, что если задана предельная ошибка δ и число наблюдений n, то можно определить отвечающую ей доверительную вероятность P по схеме

$$\delta \to z = \frac{\delta\cdot\sqrt{n}}{\sigma} \to \Phi_0(z) \to P_{дов} = 2\Phi_0(z). \qquad (3.4.9)$$

Если же заданы δ и P, то можно определить необходимый для этого объем выборки по схеме

$$P_{дов} = 2\Phi_0(z) \to z \to n = \frac{z^2\sigma^2}{\delta^2}. \qquad (3.4.10)$$

Таблица 8

f	Уровень значимости, %								
	25	10	5	2,5	1	0,5	0,25	0,1	0,05
1	1,000 0	3,077 7	6,313 8	12,706 2	31,820 5	63,656 7	127,321 3	318,308 8	636,619 2
2	0,816 5	1,885 6	2,920 0	4,302 7	6,964 6	9,924 8	14,089 0	22,327 1	31,599 1
3	0,764 9	1,637 7	2,353 4	3,182 4	4,540 7	5,840,9	7,453 3	10,214 5	12,924 0

Модифицируем использованный ранее пример к первой из рассмотренных задач. По выборке объема $n=36$ определено выборочное среднее значение урожайности пшеницы $\bar{d}=28$ ц/га. Среднее квадратическое отклонение $\sigma=1{,}8$ ц/га. Какой доверительной вероятности отвечает точность определения генерального среднего $\delta=0{,}2$ ц/га, иными словами, если сделать утверждение о том, что средняя урожайность данного сорта пшеницы равна $28\pm0{,}2$ ц/га, то какова надежность (доверительная вероятность) такого утверждения?

По схеме (3.4.9) находим

$$z=\frac{0{,}2\cdot 36}{1{,}8}=0{,}67\ .\ \Phi_0(0{,}67)=0{,}2486(\text{см. [16—18]})\ .$$

Тогда $P=2\cdot\Phi_0\ (0{,}67)=0{,}4972=49{,}72\approx 50\%$.

Таким образом, только у 50% выборок средняя урожайность будет находиться в интервале $27{,}8\div28{,}2$ ц/га. У второй половины выборок средние значения урожайности будут находиться слева или справа (поровну, по 25%) от указанных границ. Навряд ли такую надежность можно считать достаточно высокой.

Если взять $\delta=0{,}588$ ц/га, тогда имеем

$$z=\frac{0{,}588\cdot\sqrt{36}}{1{,}8}=1{,}96\cdot\Phi_0(1{,}96)=0{,}475$$

$P=2\cdot\Phi_0(1{,}96)=0{,}95=95\%$,

то есть с доверительной вероятностью 95% можно утверждать, что среднее значение урожайности пшеницы этого сорта находится внутри интервала $27{,}4\div28{,}6$ ц/га.

Определим теперь по данным этого же примера объем выборки n, при котором указанная выше предельная ошибка $\delta=0{,}2$ ц/га гарантируется с доверительной вероятностью $P=95\%$. По (3.4.10) находим

$$n=\frac{1{,}96^2\cdot1{,}8^2}{0{,}2^2}=311{,}1696\approx311\ .$$

Приводимые ниже программы П 7/34 и П 8/34 реализуют рассмотренные методы интервальных оценок для среднего.

Пример. Требуется найти доверительный интервал ($P_{\text{дов}}=95\%$) для среднего по выборке случайной вели-

чины X с нормальным законом распределения, опираясь на следующие результаты титрования [25, с. 128]: X, см³: 76,48; 76,43; 76,20; 76,45; 76,25; 76,48; 76,48; 76,60.

Решение. По таблице [17, 18] определяем значение квантиля t-распределения Стьюдента $t_{7;\,0,975}=2,36$. Согласно инструкции к программе П 8/34 после ввода программы и перехода в режим автоматической работы вводим $t_{f;\,1-\alpha/2}$ в регистр 8 и начинаем обработку данных: 2,36 П 8 76,48 С/П и т. д. По окончании работы программы имеем следующие результаты:

$\bar{X}=76,54625$ (Р2), $s^2=7,9142857\cdot10^{-2}$ (Р3),
$\delta=2,3473231\cdot10^{-1}$ (Р4), $\bar{X}_{min}=\bar{X}-\delta=76,311518$ (Р6),
$\bar{X}_{max}=\bar{X}+\delta=76,780982$ (Р7); $t_{7;\,0,975}=2,36$ (Р8),
$n=8$ (Р9).

Приводимая ниже программа П 9/34 позволяет по объему выборки n, дисперсии σ^2 и заданной точности оценки среднего значения δ определить доверительную вероятность оценки среднего с такой точностью.

Пример. Измерение длины 50 случайно отобранных после изготовления деталей дало $\bar{X}=10$ см [40, с. 170]. Известно, что генеральная дисперсия $\sigma^2=0,09$ см². Определить, с какой доверительной вероятностью можно гарантировать ошибку среднего, не превышающую 0,05 см.

Решение. Используя программу П 9/34, рассчитаем значение $P_{дов}$ по известным значениям объема выборки $n=50$, дисперсии $\sigma^2=0,09$ см² и $\delta=0,05$ см. После ввода П9/34 и выполнения п. 2 и 3 инструкции к П 9/34 вводим исходные данные задачи: 0,05↑0,09 F√↑(т. к. 0,09 — это дисперсия σ^2, а ввести надо σ, то поэтому после набора 0,09 извлекаем квадратный корень F√) 50 С/П. Результат по окончании счета $7,614071\cdot10^{-1}$, т. е. $P_{дов}\approx76\%$.

Пример. По условиям предыдущего примера определить $P_{дов}$, если объем выборки увеличен вдвое.

Решение. Согласно п. 6 инструкции к П 9/34 вводим исходные данные: 0,05↑0,09 F√↑100 С/П. Результат $9,044193\cdot10^{-1}$, т. е. $P_{дов}\approx90\%$. Ниже приведена программа 10/34, которая решает задачу нахождения минимального объема выборки, при котором гарантируется заданная точность с заданной доверительной вероятностью при оценке среднего значения и известном σ.

Пример. Определить по данным предыдущего примера объем выборки n, при котором предельная ошибка $\delta = 0{,}05$ гарантируется с вероятностью $P_{дов} = 0{,}95$.

Решение. После набора программы П 10/34 и ввода вспомогательных констант по п. 2 вводим исходные данные задачи: 0,09F√ П5 0,05 П6 0,95 В/О С/П.

Результат 138,41562, т. е. $n \approx 138$. Определим значение z: ИП7 ($z = 1{,}9608363$), т. е. $z \approx 1{,}96$.

Пример. Определить при тех же условиях объем выборки n, при котором гарантируется предельная ошибка $\delta = 0{,}05$, если в результате комплекса мероприятий удалось уменьшить дисперсию σ^2 до значения 0,01 см^2.

Решение. Согласно п. 5 инструкции вводим данные новой задачи: 0,01 F√ П5 0,05 П6 0,95 В/О С/П. Результат 15,379516, т. е. $n \approx 15$.

3.4.4. Понятие о толерантном интервале

Довольно часто о качестве разрабатываемой или выпускаемой продукции судят по тому, какая доля изделий заключена между двумя пределами, которые составляют допустимые границы изменчивости признака X. Например, два станка-автомата выпускают детали с одинаковым средним размером $\bar{X} = 10$ см. Однако для первого станка измерения показали, что 95% всех деталей имеют размеры, находящиеся внутри интервала $9{,}9 \div 10{,}1$ см, а для второго станка те же 95% выпускаемых деталей имеют размеры внутри интервала $9{,}95 \div 10{,}05$ см. Ясно, что второй станок выпускает детали более высокого качества по данному признаку.

Если нам известно значение генерального среднего μ и значение генеральной дисперсии σ^2, а также известно, что распределение вероятностей подчиняется нормальному закону, то установить границы для определенной доли всей генеральной совокупности не представляет труда.

Однако на практике чаще всего мы имеем дело с выборочными значениями как среднего, так и дисперсии. В этом случае подобные границы носят название толерантных границ, а интервал соответственно называют толерантным интервалом. В случае нормального распределения вероятностей признака X в генеральной

совокупности толерантные границы определяются по выборочным характеристикам $\bar{X}$ и s по следующим формулам:

$$X_{нижн}=\bar{X}-k\cdot s; \quad X_{верх}=\bar{X}+k\cdot s. \qquad (3.4.11)$$

Величина коэффициента k зависит от задаваемой статистической надежности (доверительной вероятности) β и задаваемой доли γ всей нормальной совокупности. В случае двухсторонних толерантных границ k является решением уравнения [17, 18]:

$$P\{P(X_{нижн}\leqslant X\leqslant X_{верх})=\gamma\}=\beta. \qquad (3.4.12)$$

Естественно, что значение k также зависит и от объема выборки n. Программа П 12/34 производит вычисление $k_{n,\,\beta,\,\gamma}$.

ЗАДАЧИ К ГЛ. 3

1. На двух делянках селекционной станции выращивали два новых сорта пшеницы. С каждой из этих делянок одновременно перед сбором урожая были взяты по 60 проб зерна. В каждой пробе находилось по 10 зерен пшеницы, взятых с одного колоса. Результаты взвешивания каждой из проб двух выборок приведены в табл. 9 и 10. Построить гистограммы, определить параметры выборок (выборка 1: $X_{min}=2$, $X_{max}=6{,}5$; выборка 2: $X_{min}=1{,}6$, $X_{max}=7$).

2. [47, с. 112]. Исследовалось распределение количества особей одного морского животного по глубине обитания. Для этого брались пробы объемом 1 м³ на разных глубинах: от 0 (у самой поверхности моря) до, примерно, 1 км, и в каждой пробе подсчитывалось число особей данного вида. При обработке первичных материалов все распределение глубин обитания было сведено в вариационный ряд с промежутком 100 м. В табл. 11 приведены результаты построения вариационного ряда приведенных к серединам интервалов X.

Используя программу П5/34, определить центральные моменты этого распределения и коэффициенты асимметрии и эксцесса.

3. Определить необходимый объем испытаний образцов с целью оценки среднего значения предела проч-

Таблица 9

4,5	3,5	3	5,5	6	4,2	2,5	2	4	5	6,2	2,2	2,8	4,7	3,7	3,5	5,6	5,1	4,3	6,3
2,3	4,8	5,2	3,8	3,4	2,5	5,7	3,9	3,7	2,6	2,1	5,3	4,9	4,3	6,4	5,8	6,5	5,7	3,7	2,5
4,0	3,6	4,4	4,5	4,2	5,4	4,4	5,2	4,2	4,3	4,9	4,4	4,7	5,3	4,8	4,5	4,8	4,8	4,6	4,7

Таблица 10

3,5	4,7	5,6	5,9	6,5	2,9	2,2	1,6	4,0	5,4	4,2	3,0	2,4	1,9	1,8	4,5	4,6	3,7	5,5	6,2
4,1	4,3	2,5	6,3	5,2	1,6	4,7	3,9	3,1	2,6	6,2	4,2	3,8	4,9	3,3	1,7	7	5,6	6,5	5,7
4,3	2,3	5,1	6,4	3,9	6,3	3,6	4,8	3,2	5,6	7,0	4,4	5,5	3,5	5,2	6,9	1,8	4,0	3,4	4,5

Таблица 11

X	50	150	250	350	450	550	650	750	850	950	1050
n	1	45	92	117	76	45	24	11	5	3	1

Таблица 12

Объем n_i	30	30	30	30	30	30	30	30	30	30
Среднее $\overline{d}_i$, мкм	11,7	12,5	11,8	11,2	12,3	12,0	11,7	12,9	12,3	12,0
Стандарт s_i, мкм	0,8	0,9	1,2	0,9	1,4	0,9	0,4	0,5	1,2	1,1
Объем n_i	30	30	30	30	30	30	30	30	30	30
Среднее $\overline{d}_i$, мкм	11,9	12,5	11,8	12,4	12,5	12,5	11,2	11,4	12,3	11,8
Стандарт s_i, мкм	1,0	0,4	0,7	0,6	0,9	0,8	0,5	0,4	0,7	0,8

ности алюминиевого сплава, если требуемая доверительная вероятность $P_{дов} = 95\%$, требуемая точность оценки $\delta = 10$ МПа, а значение $\sigma = 30$ МПа.

4. В цехе завода осуществляется освоение нового штамповочного автомата. Технолог взял из деталей, отштампованных на этом автомате, три выборки объемом 90 деталей. Выборка 1 взята из партии деталей, изготовленных в 1-ю смену, выборка 2 соответствует 2-й смене и т. д.

Требуется провести сравнительный анализ трех гистограмм и проверить полученные при этом выводы с помощью гистограммы суммарного распределения. Ниже приведены значения веса деталей в граммах по каждой выборке. Для построения гистограмм использовать программу П1/34 с числом интервалов, равным 7.

Выборка 1

51; 43; 40; 49; 54; 47,5; 45; 41; 43,5; 45,5; 46,5; 49,5; 53; 51,5; 43; 51; 45,5; 47; 49; 45; 49; 47; 45,5; 49,5; 48; 47,5; 49; 47; 47,5; 47.

Выборка 2

49; 49,5; 49; 51; 49,5; 50; 51,5; 47,5; 49; 51; 48,5; 51,5; 55; 53,5; 45; 53; 47,5; 51; 47; 53; 45; 42; 51; 56; 49,5; 47; 43; 45,5; 47,5; 48.

Выборка 3

45; 48; 46; 50; 52,5; 59; 56; 54; 48,5; 58; 50,5; 51,5; 56,5; 54,5; 48; 56; 50,5; 52; 54; 50; 54; 52; 50,5; 53; 54,5 52,5; 54; 52; 52,5; 52.

5. При разработке технологии химического никелирования на детали сложной конфигурации были исследованы 20 выборок. Каждая выборка содержала данные по 30 измерениям толщины осажденного слоя никель-фосфорного сплава. Результаты измерений (средняя толщина и стандарт) приведены в табл. 12. Требуется на основании этих данных получить оценку средней толщины покрытия, стандартного отклонения и ошибки среднего значения (стандартного отклонения среднего).

6. По аналогии с условиями задач 3 и 4 составьте и решите задачи по собственной тематике исследований. Проведите сравнительный анализ полученных гистограмм.

РЕШЕНИЯ

1. Согласно формуле Штюргеса определяем количество интервалов $k=7$ для построения гистограмм. После набора программы П1/34 вводим $X_{min}=2$ для первой выборки и X_{max} : 2↑6,5 В/О С/П (—4,5). Вводим k : 7 С/П (0). Приступаем к вводу первой выборки: 4,5 С/П (1) 3,5 С/П (1) 3 С/П (1) 5,5 С/П (1) и т. д. По окончании ввода выполняем п. 6—11: БП 33 С/П (60) 2 С/П ($1{,}3333333\cdot10^{-1}=p_1$)
F, ($8=m_1$) F, ($2{,}6428572=X_{12}$) С/П ($3{,}3333333\cdot10^{-2}=p_2$)
F, ($2=m_2$) F, ($3{,}2857144=X_{23}$) С/П ($1{,}5\cdot10^{-1}=p_3$)
F, ($9=m_3$) F, ($3{,}9285716=X_{34}$) С/П ($2{,}3333333\cdot10^{-1}=p_4$)
F, ($14=m_4$) F, ($4{,}5714288=X_{45}$) С/П ($2{,}3333333\cdot10^{-1}=p_5$)
F, ($14=m_5$) F, ($5{,}214286=X_{56}$) С/П ($1{,}3333333\cdot10^{-1}=p_6$)
F, ($8=m_6$) F, ($5{,}8571432=X_{67}$) С/П ($8{,}3333333\cdot10^{-2}=p_7$)
F, ($5=m_7$) F, ($6{,}5000004=X_{78}$). Таким образом, получены все исходные данные для построения гистограммы по первой выборке.

Аналогично обработав вторую выборку, получаем следующие результаты: БП 33 С/П (60), 1,6 С/П ($1{,}3333333\cdot10^{-1}=p_1$) F, ($8=m_1$) F, ($2{,}371429=X_{12}$) С/П ($1\cdot10^{-1}=p_2$) F, ($6=m_2$) F, ($3{,}1428577=X_{23}$) С/П ($1{,}6666666\cdot10^{-1}=p_3$), ($10=m_3$) F, ($3{,}9142864=X_{34}$) С/П ($1{,}8333333\cdot10^{-1}=p_4$) F, ($11=m_4$) F, ($4{,}8657151=X_{45}$) С/П ($1{,}5\cdot10^{-1}=p_5$) F, ($9=m_5$) F, ($5{,}4571438=X_{56}$) С/П ($1{,}3333333\cdot10^{-1}=p_6$) F, ($8=m_6$) F, ($6{,}2285725=X_{67}$) С/П ($1{,}3333333\cdot10^{-1}=p_7$) F, ($8=m_7$) F, ($7{,}0000012=X_{78}$).

Имеем исходные данные для построения второй гистограммы.

Вводим программу П3/34 и начинаем обработку первой выборки: В/О 4,5 С/П (1) 3,5 С/П (2) 3 С/П (3) 5,5 С/П (4) и т. д. После ввода всей выборки на индикаторе число 60 — объем введенной выборки. Вычисляем оценки статистик: $\vec{Ш}$ С/П ($26{,}15843=v$, %) ИПА ($4{,}3566666=\bar{X}$) ИПВ ($1{,}1396356=s$) ИПД ($1{,}4712632\cdot10^{-1}=s_{\bar{x}}$).

Аналогично обрабатываем вторую выборку. $\vec{Ш}$ С/П ($35{,}008316=v$, %) ИПА ($4{,}3=\bar{X}$) ИПВ ($1{,}5053576=s$) ИПД ($1{,}9434083\cdot10^{-1}=s_{\bar{x}}$).

Поскольку мы имеем для обеих выборок все необходимые данные для построения доверительного интервала

для среднего ($\bar{X}$, $s_{\bar{x}}$ n), то строим 95%-ные доверительные интервалы. По таблице [17, 18] находим значение $t_{0,025;\,59}=2{,}0$. Тогда $\delta=2\cdot s_{\bar{x}}=2\cdot 1{,}4712632\cdot 10^{-1}\approx 0{,}294$. Таким образом, для первой выборки имеем границы доверительного интервала для $\bar{X}$: $4{,}357\pm 0{,}294$, т. е. $\bar{X}_{min}=4{,}063$, $\bar{X}_{max}=4{,}651$. Аналогично для второй выборки: $\delta=2\cdot s_{\bar{x}}=2\cdot 1{,}9434083\cdot 10^{-1}\approx 0{,}388$; $\bar{X}\pm\delta=4{,}3\pm 0{,}388$; $\bar{X}_{min}=3{,}912$, $\bar{X}_{max}=4{,}688$.

Таким образом, доверительные интервалы для средних обеих выборок построены. Как видно по границам этих интервалов, они (интервалы) достаточно сильно перекрываются. Что может означать такое наложение этих двух интервалов? Что можно сказать об урожайности этих двух сортов на основе полученных результатов обработки?

2. После ввода программы П5/34 и выполнения п. 2 приступаем к вводу таблицы: 50↑1 С/П (50) 150↑45 С/П (150) и т. д. По окончании ввода таблицы выполняем БП 27 С/П ($385=M_1$) С/П ($175833{,}33=M_2$) С/П ($92976785=M_3$) С/П ($5{,}5728154\cdot 10^{10}=M_4$). Далее вносим длину интервалов h и вычисляем центральные моменты: 100 БП 36 С/П ($385=m_1=\bar{X}$) С/П ($26774{,}997=m_2=s^2$) С/П ($4022539=m_3$) С/П ($2{,}874198\cdot 10^9=m_4$). Вычисляем коэффициенты асимметрии и эксцесса: ИП3 ИП4↑FX²+XF√÷ПА ($9{,}1813411\cdot 10^{-1}$А*) ИП2 ИП4 FX²÷÷3—ПВ ($1{,}0092009=$E*). Таким образом, исследуемое распределение особей морского животного по глубине имеет положительную асимметрию и положительный эксцесс, т. е. распределение имеет смещение моды в сторону меньших глубин обитания и острую вершину.

3. После набора программы П 10/34 и ввода вспомогательных констант вводим исходные данные в ПМК: 30 П5 10 П6 0,95. Вычисляем требуемый объем выборки: В/О С/П ($34{,}603911\approx 35=n$). Таким образом, для удовлетворения требуемых условий достаточно будет проведения испытания 35 образцов.

4. Результаты применения П1/34 к первой выборке (принимаем $X_{min}=40$; $X_{max}=55$; $k=7$); $n=30$;

$p_1=6{,}6666666\cdot 10^{-2}$; $m_1=2$; $X_{12}=42{,}142857$;

$p_2=1\cdot 10^{-1}$; $m_2=3$; $X_{23}=44{,}285714$;

$p_3 = 1{,}6666666 \cdot 10^{-1}$; $m_3 = 5$; $X_{34} = 46{,}428571$;
$p_4 = 3 \cdot 10^{-1}$; $m_4 = 9$; $X_{45} = 48{,}571428$;
$p_5 = 2 \cdot 10^{-1}$, $m_5 = 6$; $X_{56} = 50{,}714285$;
$p_6 = 1 \cdot 10^{-1}$, $m_6 = 3$; $X_{67} = 52{,}857142$;
$p_7 = 6{,}6666666 \cdot 10^{-2}$; $m_7 = 2$; $X_{78} = X_{max} = 54{,}999999$.

Результаты второй выборки идентифицируйте сами, обработав выборку по П1/34 с $X_{min} = 41$; $X_{max} = 56$; $k = 7$: 2; 2; 3; 10; 8; 3; 2; $6{,}6666666 \cdot 10^{-2}$; $6{,}6666666 \cdot 10^{-2}$; $1 \cdot 10^{-3}$; $3{,}3333333 \cdot 10^{-1}$; $2{,}6666666 \cdot 10^{-1}$; $1 \cdot 10^{-1}$; $6{,}6666666 \cdot 10^{-1}$; 55,999999; 53,837142; 51,714285; 49,571428; 47,428571; 45,285714; 43,142857; 30.

Аналогично идентифицируйте и результаты по второй выборке $X_{min} = 44$; $X_{max} = 60$, $K = 7$):
$6{,}6666666 \cdot 10^{-2}$; $1 \cdot 10^{-1}$; $1{,}6666666 \cdot 10^{-1}$; $3 \cdot 10^{-1}$; $2 \cdot 10^{-1}$; $1 \cdot 10^{-1}$; $6{,}6666666 \cdot 10^{-2}$; 30; 2; 3; 5; 9; 6; 3; 2; 46,285715; 48,57143; 50,857145; 53,14286; 55,428575; 57,71429; 60,000005.

Совместная обработка трех выборок по П1/34 с $X_{min} = 40$, $X_{max} = 60$ и $k = 7$ дает следующие результаты:

$n = 90$; $p_1 = 3{,}3333333 \cdot 10^{-2}$; $m_1 = 3$; $X_{12} = 42{,}857143$;

$p_2 = 1{,}4444444 \cdot 10^{-1}$; $m_2 = 13$; $X_{23} = 45{,}714286$;

$p_3 = 2{,}2222222 \cdot 10^{-1}$; $m_3 = 20$; $X_{34} = 48{,}571429$;

$p_4 = 2{,}6666666 \cdot 10^{-1}$; $m_4 = 24$; $X_{45} = 51{,}428572$;

$p_5 = 2{,}3333333 \cdot 10^{-1}$; $m_5 = 21$; $X_{56} = 54{,}285715$;

$p_6 = 7{,}7777777 \cdot 10^{-2}$; $m_6 = 7$; $X_{67} = 57{,}142858$;

$p_7 = 2{,}2222222 \cdot 10^{-2}$; $m_7 = 2$; $X_{78} = X_{max} = 60{,}000001$.

Построив на основе этих результатов четыре графических гистограммы, можно увидеть, что положение моды распределения смещается вправо от первой выборки к третьей. По этой причине суммарная гистограмма характеризуется плосковершинностью ($m_3 = 20$; $m_4 = 24$; $m_5 = 21$).

5. С помощью программы П 4/34 получаем следующие результаты: $\bar{\bar{X}} = 12{,}010526$ мкм; $s = 9{,}5684932 \cdot 10^{-1}$ мкм; $v = 7{,}9667561\,\%$; $s_x = 4{,}0113211 \cdot 10^{-2}$ мкм.

ГЛАВА 4

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Даже самая сильная вера должна быть подкреплена статистическим критерием.

В. Феллер

4.1. Понятие о статистических гипотезах

На основании выборочных оценок, методы нахождения которых рассмотрены в главе 3, в конкретных исследованиях и разработках необходимо принимать столь же конкретные решения. Такие решения могут выбираться из двух либо большего количества альтернативных решений. Примеры таких ситуаций приводились уже в главе 3. Напомним наиболее типичные из них. В результате разработанной технологии производства изделия параметр качества X_1 должен иметь среднее значение $\bar{X}_1 = A$. После проведения установочной партии в цехе завода получена выборка изделий объемом n. Среднее значение параметра качества в этой выборке изделий равно значению A_1. Можно ли считать, что отклонение A_1 от A случайно и незначимо? Аналогичный вопрос может быть поставлен и в отношении других параметров распределения, а также и о самом виде распределения. Например, на основе полученной выборки требуется ответить на вопрос: можно ли считать, что распределение нормально? Такие вопросы, требующие ответа и, следовательно, определенных, конкретных решений и действий с научными и технико-экономическими последствиями, возникают в исследовательской и производственной практике практически ежедневно. В таких случаях принято говорить о проверке статистических гипотез.

История проверки статистических гипотез восходит к XVIII веку. Самый первый пример проверки такой статистической гипотезы описан в работе, датированной 1710 г. и написанной Дж. Арбутнотом [19]. Проанализировав систематические наблюдения в течение 82 лет за рождением детей, он отметил, что мальчиков рожда-

ется больше, чем девочек. Эти данные опровергали гипотезу о том, что рождение мужчин и женщин равновероятно (с вероятностью 1/2).

Важно различать научную и статистическую гипотезы. Не все научные гипотезы требуется проверять статистически и, наоборот, не все статистические гипотезы обязательно должны представлять научный интерес.

Приведем другой пример, ставший хрестоматийным [15]. До работ французского ученого Луи Пастера, основоположника современной микробиологии и иммунологии, большинство ученых придерживались теории самозарождения микроорганизмов. Для подтверждения этого обычно приводили факт «самозарождения» белых личинок в тухлом мясе. Для проверки этой теории Пастер выставил в одном и том же месте близкие по качеству и величине куски начинавшего портиться мяса. Одни из кусков были защищены от мух, на другие же мухи могли садиться. Личинки появились на тех кусках, к которым мухи имели доступ. На тех же кусках, которые были защищены от мух, личинок не оказалось. В свете этих фактов гипотеза о самозарождении личинок была отвергнута. Полученные тогда результаты для нас сейчас очевидны и для придания достоверности этим выводам нет необходимости использовать методы статистики, поскольку данная научная гипотеза и не нуждается в них. Однако такая очевидность выводов и результатов экспериментов имеется не всегда и в тех же биологических науках. Достаточно сказать, что в связи с анализом именно неочевидных результатов в биологических экспериментах такие видные основатели современной прикладной статистики, как Р. Фишер, К. Пирсон, Дж. Снедекор и др. разработали многие из методов, которыми мы сейчас пользуемся. При изучении таких неочевидных результатов гипотезы формулируются в виде статистических гипотез.

Итак, что же такое статистическая гипотеза? Статистической называют гипотезу (предположение) о виде неизвестного распределения или о параметрах известных распределений. Процедура обоснованного сопоставления высказанной статистической гипотезы с имеющимися в нашем распоряжении выборочными данными $X_1, X_2, \ldots, X_n$ осуществляется с помощью того или иного статистического критерия

и называется проверкой статистических гипотез.

Например, статистическими будут следующие гипотезы:

1) Дисперсии двух генеральных совокупностей равны.

2) Средние значения двух генеральных совокупностей равны.

3) Генеральная совокупность подчиняется нормальному закону распределения.

В первых двух гипотезах сделано предположение о параметрах двух известных распределений, в третьей — о виде неизвестного распределения.

4.1.1. Ошибки первого и второго рода

В практике статистического анализа чаще всего приходится иметь дело с двумя конкурирующими гипотезами: нулевой гипотезой, обозначаемой H_0, и альтернативной гипотезой, обозначаемой H_1. Нулевая гипотеза — это гипотеза, подлежащая проверке, и если мы отвергаем нулевую гипотезу как неподходящую в каком-то статистическом смысле, то мы принимаем альтернативную гипотезу. Поскольку проверка проводится статистическими методами, то гипотезу называют статистической. Учитывая то, что проверка гипотезы производится на основе выборочной информации, абсолютную правильность решения гарантировать нельзя. Поэтому не исключена вероятность того, что в результате проверки гипотезы может быть допущена ошибка и принято неверное решение. В этом случае могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза.

Ошибка второго рода состоит в том, что будет принята неправильная гипотеза.

Смысл обеих этих ошибок совершенно различен, так же как и последствия совершения ошибки первого и второго рода.

В свою очередь, и правильное решение также может быть принято в двух случаях:

1) Принимается нулевая гипотеза H_0, когда она действительно верна.

2) Отвергается нулевая гипотеза H_0, когда она действительно неверна.

Вероятность ошибки первого рода принято обозначать через α и называть уровнем значимости. Величину α наиболее часто принимают равной 5 или 1%. Значение задается самим исследователем до начала процедуры статистической проверки. Если $\alpha=5\%$, то это означает, что в пяти случаях из ста мы рискуем допустить ошибку первого рода, т. е. отвергнуть правильную гипотезу. Нулевой эту гипотезу называют по той причине, что часто эта гипотеза заключается в предположении об отсутствии значимого расхождения (нулевом расхождении) между неизвестным параметром генеральной совокупности и заданной величиной. Содержание гипотезы записывается через двоеточие, например $H_0 : X=\mu_0$.

Совершенно исключить ошибки первого и второго рода при статистической проверке гипотез невозможно. Для минимизации потерь от подобных ошибок, видимо, целесообразно уменьшить обе ошибки одновременно. Однако такое требование противоречиво, т. к. эти ошибки являются конкурирующими и уменьшение вероятностей одной из них влечет за собой увеличение вероятности другой. В подавляющем большинстве случаев такое требование выполнимо только при увеличении объема выборок.

С помощью методов статистической проверки гипотез решаются два основных класса задач. К первому из них относятся задачи сравнения: сравнения выборочных характеристик (среднего, дисперсии и т. п.) с заданным нормативным, номинальным значением, а также сравнения выборочных характеристик двух выборок между собой. Ко второму классу задач можно отнести задачи о проверке принадлежности выборки генеральной совокупности с определенным законом распределения и ряд других аналогичных задач.

Методы статистической проверки гипотез широко используются практически во всех видах статистического анализа, в частности, в корреляционном, регрессионном дисперсионном анализах и в методах многомерной статистики. В данной главе мы кратко рассмотрим ряд задач первого и второго класса, возникающих в исследовательской и производственной практике.

4.2. Проверка статистических гипотез о равенстве дисперсий

Рассмотрим следующую производственную ситуацию. Изготовлены два опытных образца терморегуляторов, которые установлены в две одинаковые установки. Ниже приведены замеры температуры в °С, производимые на каждой установке через 10 мин.

Установка 1: 97,8; 98,9; 101,2; 98,8; 102,0; 100,5; 100,9; 99,9; 100,8; 99,9; 99,0; 99,1; 99,9 (13 замеров).

Установка 2: 97,2; 97,4; 96,8; 97,9; 99,9; 97,2; 100,5; 98,2; 98,3; 97,5 (10 замеров).

Можно ли на основе этих результатов сказать, что один из образцов терморегулятора обеспечивает более стабильное поддержание заданной температуры?

Согласно изложенной выше схеме сформулируем нулевую гипотезу H_0:

$$H_0 : \sigma_1^2 = \sigma_2^2,$$

где σ_1^2 и σ_2^2 — дисперсии температуры в установках 1 и 2. Поскольку дисперсии σ_1^2 и σ_2^2 неизвестны, то проверку выдвинутой гипотезы H_0 осуществляют на основе сопоставления выборочных дисперсий s_1^2 и s_2^2. Очевидно, что если отношение s_1^2/s_2^2 близко к 1, то это является аргументом в пользу принятия нулевой гипотезы. Если же это отношение значительно отличается от 1, то это аргумент в пользу отклонения нулевой гипотезы. Зададимся уровнем значимости $\alpha = 5\%$. Статистическим критерием в данном случае является отношение s_1^2/s_2^2 при $s_1^2 > s_2^2$.

Английский статистик Р. Фишер исследовал распределение величины $z = 0{,}5 \ln(s_1^2/s_2^2)$, где s_1^2 и s_2^2 — выборочные оценки дисперсии. Несколько позднее работ Р. Фишера американский статистик Д. У. Снедекор предложил перейти к рассмотрению непосредственно отношения $F = s_1^2/s_2^2$, которое называется F-отношением. F-распределение Фишера—Снедекора (иногда это распределение называют распределением Фишера) зависит только от числа степеней свободы f_1 и f_2, по которым подсчитываются выборочные дисперсии s_1^2 и s_2^2, т. е. $f_1 = n_1 - 1$, а $f_2 = n_2 - 1$, где n_1 и n_2 — объемы первой и второй выборок. При нахождении F-отношения берется отношение большей дисперсии к меньшей.

Дальнейшая процедура статистической проверки выдвинутой нулевой гипотезы H_0 зависит от того, какую альтернативную ей гипотезу H_1 мы выдвигаем. Именно от этого зависит определение критического значения F-отношения при заданном уровне значимости α.

Рассмотрим первый случай:

$$H_0 : \sigma_1^2 = \sigma_2^2;$$

$$H_1 : \sigma_1^2 > \sigma_2^2.$$

В этом случае строят одностороннюю, а именно правостороннюю критическую область, исходя из следующего требования: вероятность попадания F-отношения (F-критерия) в эту область, в предположении справедливости нулевой гипотезы H_0, должна быть равна принятому значению α:

$$P\{F > F_{кр}(\alpha, f_1, f_2)\} = \alpha. \qquad (4.2.1)$$

Критическую точку $F_{кр}(\alpha, f_1, f_2)$ находят по таблице значений $F(f_1, f_2)$ для разных α [17, 18]. Тогда правосторонняя область будет определяться неравенством

$$F > F_{кр},$$

т. е. если $F = (s_1^2/s_2^2) > F_{кр}$, то нулевую гипотезу H_0 необходимо отвергнуть с риском в 5% (при $\alpha = 5\%$).

Соответственно область принятия нулевой гипотезы H_0 определяется другим неравенством

$$F < F_{кр},$$

т. е. если $F = (s_1^2/s_2^2) < F_{кр}$, то нулевую гипотезу следует принять.

Для нашего случая $s_1^2 = 1{,}69$, а $s_2^2 = 1{,}44$. Первоначально проверим одностороннюю гипотезу, т. е.

$$H_0 : \sigma_1^2 = \sigma_2^2;$$

$$H_1 : \sigma_1^2 > \sigma_2^2.$$

В нашем случае F-отношение равно

$$F = \frac{1{,}69}{1{,}44} = 1{,}17 .$$

Критическая точка $F_{(0{,}05;\ 12{,}9)} = 3{,}1$. Поскольку $F < F_{кр}$, то нулевая гипотеза принимается, и у нас нет оснований считать, что второй терморегулятор обеспечивает более точное поддержание температуры.

Нередко альтернативную гипотезу формулируют иначе: $H_1 : \sigma_1^2 \neq \sigma_2^2$. Следует иметь в виду, что эти две альтернативные гипотезы существенно отличаются друг от друга. В чем их отличие? В первом случае рассматривается гипотеза о том, что дисперсия $\sigma_1^2 > \sigma_2^2$, а во втором — $\sigma_1^2 \neq \sigma_2^2$. Это неравенство можно представить иначе, а именно в виде двух неравенств: $\sigma_1^2 > \sigma_2^2$ либо $\sigma_1^2 < \sigma_2^2$. В этом случае строится двухсторонняя критическая область. При построении такой области исходят из требования, чтобы вероятность попадания F-критерия в такую область, при условии справедливости нулевой гипотезы, была равна заданному уровню значимости α.

Было установлено, что максимальная мощность F-критерия, т. е. вероятность попадания этого критерия в критическую область, когда действительно верна альтернативная гипотеза H_1, достигается в том случае, если вероятность попадания F-критерия в каждый из двух интервалов равна $\alpha/2$. Обозначив через F_1 левую границу критической области 1 и через F_2 — правую, будем иметь

$$P(F < F_1) = \frac{\alpha}{2}; \quad P(F > F_2) = \frac{\alpha}{2}. \qquad (4.2.2)$$

В этом случае область принятия нулевой гипотезы H_0 будет отвечать значениям:

$$F_1 < F < F_2. \qquad (4.2.3)$$

Правая критическая точка $F_2 = F_{кр}\left(\frac{\alpha}{2};\ f_1;\ f_2\right)$ находится по таблицам F-распределения при уровне значимости $\alpha/2$ и степенях свободы f_1 и f_2.

Левых же точек в этой таблице нет и найти F_1 по этим таблицам невозможно. Однако этого и не требуется, т. к. достаточно определения только F_2.

Определим критическую точку $F_{кр(0,025;\ 12,9)} = 3,9$. Поскольку $F < F_{кр} = 3,9$, то и в этом случае у нас нет оснований отвергнуть нулевую гипотезу. Это означает, что с доверительной вероятностью $P_{дов} = 95\%$ можно утверждать, что оба образца терморегулятора обеспечивают одинаковую точность регулирования температуры в данных установках.

Метод статистической проверки гипотез применяется и в том случае, когда требуется проверить равенство

выборочной s^2 и гипотетической генеральной дисперсии σ_0^2. В этом случае нулевую гипотезу H_0 можно записать следующим образом:

$$H_0 : \sigma^2 = \sigma_0^2.$$

В практической деятельности такая ситуация возникает в тех случаях, когда на основании выборочных наблюдений необходимо проверить точность измерительных приборов, инструментов, исследовать устойчивость технологического процесса. Например, известна точность измерений концентрации носителей тока в полупроводнике, выполняемых с помощью установки измерения с использованием C/V метода, и необходимо проверить, соответствует ли фактическая точность допустимой.

Для проверки такой нулевой гипотезы используется статистика $\chi^2 = \frac{(n-1)\cdot s^2}{\sigma_0^2}$, которая имеет χ^2-распределение Пирсона с $f = n-1$ степенями свободы.

В зависимости от вида альтернативной, конкурирующей гипотезы критическая область строится по разному. В первом случае

$$H_0 : \sigma^2 = \sigma_0^2, \qquad (4.2.4)$$

$$H_1 : \sigma^2 > \sigma_0^2.$$

В этом случае строится правосторонняя критическая область, исходя из требования

$$P\{\chi^2 > \chi^2_{кр}, (\alpha; f)\} = \alpha.$$

Критическую точку $\chi^2_{кр}$ $(\alpha; f)$ находят по таблице χ^2-распределения [17]. Тогда области принятия нулевой гипотезы отвечает неравенство

$$\chi^2 < \chi^2_{кр}(\alpha; f), \qquad (4.2.5)$$

и наоборот, при $\chi^2 > \chi^2_{кр}(\alpha; f)$ нулевую гипотезу отвергают.

Во втором случае имеем

$$H_0 : \sigma^2 = \sigma_0^2,$$

$$H_1 : \sigma^2 \neq \sigma_0^2. \qquad (4.2.6)$$

В этом случае строится двухсторонняя критическая область. В таблицах χ^2-распределения указаны лишь

«правые» критические точки. Учитывая, что одновременно не может быть $\sigma^2 > \sigma_0^2$ и $\sigma^2 < \sigma_0^2$, т. к. эти события противоположны, можно показать, что левая критическая точка равна

$$\chi^2_{кр}\left(1-\frac{\alpha}{2};\ f\right),\ \text{а правая}\ \chi^2_{кр}\left(\frac{\alpha}{2},\ f\right).$$

Рассмотрим пример. При проведении измерения концентрации электронов на структурах арсенида галлия проверялась точность измерений. Согласно паспорту на установку она должна обеспечивать в диапазоне концентраций $1\cdot10^{15} \div 1\cdot10^{16}$ см$^{-3}$ дисперсию измерений $1\cdot10^{28}$ см$^{-6}$. После проведения 31 измерения была найдена выборочная дисперсия $s^2 = 3{,}6\cdot10^{27}$ см$^{-6}$. Требуется установить, соответствует ли фактическая точность паспортному значению при уровне значимости $\alpha = 5\%$?

Сформулируем статистические гипотезы. Относительно нулевой гипотезы H_0 сомнений нет:

$$H_0 : \sigma^2 = \sigma_0^2.$$

Однако неясно, какую выбрать альтернативную гипотезу H_1. Здравый смысл подсказывает, что для нас очень важно, чтобы не было случая $\sigma^2 > \sigma_0^2$, т. к. в этом случае фактическая точность будет меньше (дисперсия больше), чем паспортная. И наоборот, если $\sigma^2 < \sigma_0^2$, то это будет означать, что фактическая точность выше паспортной (благоприятная ситуация). Таким образом, в качестве альтернативной гипотезы, очевидно, имеет смысл выбрать неблагоприятную ситуацию

$$H_1 : \sigma^2 > \sigma_0^2.$$

В этом случае строится односторонняя критическая область с $\chi^2_{кр}(\alpha;\ f)$. По таблице [17, 18] находим значение $\chi^2(0{,}05;\ 30) = 43{,}77$.

Находим расчетное (фактическое) значение критерия χ^2:

$$\chi^2 = \frac{30\cdot3{,}6\cdot10^{27}\ см^{-6}}{1\cdot10^{28}\ см^{-6}} = 10{,}8\,. \qquad (4.2.7)$$

Поскольку $\chi^2 < \chi^2_{кр}$, то нулевая гипотеза H_0 принимается, т. е. установка обеспечивает паспортную точность измерений.

4.3. Проверка статистических гипотез о равенстве средних

Пожалуй, наиболее часто в исследовательской и производственной практике приходится использовать метод статистических гипотез при проверке гипотез о равенстве средних. Вот наиболее типичные из подобных примеров.

1. Внесено изменение (производственная проба, предварительное извещение, рацпредложение и т. д.) в технологию изготовления изделия. Цель изменения — увеличение значения параметра качества X. По двум выборкам N_1 и N_2, взятых из изделий до и после внесения изменения либо изготовленных параллельно, нужно проверить гипотезу о том, что среднее значение параметра $\bar{X}_2 > \bar{X}_1$. Если гипотеза будет принята, то это аргумент в пользу предложенного изменения технологии.

2. Внесено изменение в технологию, которое дает положительный экономический эффект за счет уменьшения материалоемкости и трудоемкости. Требуется проверить гипотезу о том, что такое изменение не повлекло за собой уменьшения среднего значения параметра качества X, т. е. $\bar{X}_1 = \bar{X}_2$.

3. Для изменения распределения параметра качества у выпускаемого изделия по группам необходимо изменить среднее значение параметра качества X у исходного материала на величину D. После того как соответствующие изменения в технологию изготовления исходного материала внесены, требуется проверить гипотезу о том, что среднее значение параметра X действительно сместилось на величину D.

Таков далеко не полный перечень ситуаций, когда требуется проверка гипотезы о средних значениях выборок. Следует, однако, иметь в виду, что в данном случае речь идет об одномерных непрерывных переменных X, а не о векторе **X**. Случаи, когда необходимо сравнение векторов средних, будут рассмотрены ниже в главах, посвященных многомерным методам прикладной статистики.

4.3.1. Сравнение средних при известных σ_1^2 и σ_2^2

Наиболее просто процедура сравнения двух средних нормальных генеральных совокупностей осуществляется, если известны их дисперсии. В этом случае на основе выборок объемов m и n, которые взяты из этих совокупностей, определяются выборочные средние $\bar{X}$ и $\bar{Y}$. Задав уровень значимости α, требуется проверить нулевую гипотезу $H_0: \bar{X}=\bar{Y}$. В этом случае, когда средние обеих совокупностей действительно равны, незначимые различия между выборочными средними $\bar{X}$ и $\bar{Y}$ случайны, т. е. объясняются случайным характером отбора выборок. В качестве статистического критерия, используемого для проверки гипотезы, используют величину z:

$$z = \frac{\bar{X} - \bar{Y}}{\sigma_{(\bar{X}-\bar{Y})}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}, \qquad (4.3.1)$$

которая является нормированной нормальной случайной величиной. В зависимости от вида альтернативной гипотезы используется односторонняя или двухсторонняя критическая область.

В первом случае имеем

$$H_0: \bar{X}=\bar{Y};$$

$$H_1: \bar{X}>\bar{Y}. \qquad (4.3.2)$$

Именно такую одностороннюю статистическую гипотезу необходимо проверять, когда нужно убедиться в том, что изменение технологического процесса должно приводить к увеличению среднего значения какого-то параметра качества выпускаемого изделия.

В этом случае используется правосторонняя критическая гипотеза. Несложно доказать, что в этом случае

$$\Phi_0(z_{кр}) = \frac{1 - 2\alpha}{2}, \qquad (4.3.3)$$

где Φ_0 — функция Лапласа. Это означает, что для того, чтобы найти границу правосторонней критической области $z_{кр}$, достаточно найти значение аргумента функции Лапласа, которому соответствует значение функции, равное $\frac{1 - 2\alpha}{2}$. Тогда область принятия нулевой гипо-

тезы определяется неравенством $z<z_{кр}$, а область ее отклонения $z>z_{кр}$.

Ниже приводится программа П 13/34, позволяющая производить сравнение двух средних нормальных генеральных совокупностей, дисперсии которых известны по результатам выборочных испытаний (наблюдений). Рассмотрим пример на применение данной программы.

Пример. В течение 2 лет после освоения нового типа генераторного диода цех выпускал диоды этого типа в двух модификациях, отличающихся технологическим маршрутом изготовления. В самом начале освоения этого типа диодов один из маршрутов обеспечивал получение более дефицитной группы, отличавшейся большей мощностью генерации. Другой маршрут позволял производить с меньшей себестоимостью диоды несколько пониженной мощности. За 2 года этот маршрут подвергся ряду изменений, которые должны были поднять мощность выпускаемых диодов. При проверке гипотезы о том, что диоды, выпускаемые по маршруту более дефицитных групп, действительно имеют среднюю мощность генерации выше, нежели диоды, выпускаемые по параллельному маршруту, но имеющими меньшую себестоимость, было известно, что дисперсия для первого маршрута (маршрут более мощных диодов) равна 3200 мВт^2, а дисперсия для второго маршрута равна 3025 мВт^2. В данном случае был применен метод статистических гипотез.

Решение. Производя случайную выборку диодов из первого маршрута, определили $\bar{P}_1=180$ мВт. Объем выборки составил 90 диодов. Аналогичным образом была взята выборка диодов объемом $n_2=50$ диодов со средней мощностью $\bar{P}_2=165$ мВт.

Сформулированные при этом гипотезы имели вид

$$H_0 : \bar{P}_1=\bar{P}_2;$$

$$H_1 : \bar{P}_1>\bar{P}_2. \qquad (4.3.4)$$

Проверка гипотез производилась при значении уровня значимости $\alpha=5\%$ (вводим в ПМК 0,05).

Согласно инструкции к П 13/34 введем исходные данные в ПМК 34 и, вычислив разность $\Delta z=1{,}106\cdot10^{-1}$, увидим, что эта разность положительная, т. е. $z_{кр}>z$. Следовательно, нулевая гипотеза принимается и нет оснований утверждать, что средняя мощность диодов

второго маршрута меньше, чем средняя мощность диодов, выпускаемых по первому маршруту. Учитывая, что себестоимость диодов, изготавливаемых по второму маршруту, меньше, чем у диодов первого маршрута, видимо, необходимо рассмотреть целесообразность ликвидации первого маршрута.

При помощи программы П 13/34 можно производить проверку и такой комбинации H_0 и H_1:

$$H_0 : \bar{X} = \bar{Y};$$

$$H_1 : \bar{X} < \bar{Y}. \qquad (4.3.5)$$

В этом случае также при $\Delta z > 0$ принимается нулевая гипотеза H_0, а при $\Delta z < 0$ H_0 отвергается. Следует отметить, что этот случай не идентичен предыдущему, когда принималось $H_1 : \bar{X} > \bar{Y}$.

Часто при проверке статистических гипотез о средних нельзя априори сделать предположение о том, какую выдвинуть альтернативную гипотезу: $H_1 : \bar{X} > \bar{Y}$ или $H_1 : \bar{X} < \bar{Y}$. В этом случае выдвигается альтернативная гипотеза $H_1 : \bar{X} \neq \bar{Y}$ и возможно применение программы П 13/34, только в дополнение к инструкции необходимо помнить, что при этом строится двухсторонняя критическая область и, следовательно, необходимо оперировать с величиной $\alpha/2$ вместо α. Для этого можно либо сразу по п. 4 инструкции к П 13/34 набрать вместо α значение $\alpha/2$, либо, набрав значение α, нажать клавиши 2÷ и только затем запустить программу на счет: В/О С/П. Как и ранее, отрицательная разность Δz означает, что необходимо отвергнуть нулевую гипотезу о равенстве средних генеральных совокупностей, и наоборот, при $\Delta z > 0$ принять нулевую гипотезу H_0.

4.3.2. Сравнение средних при неизвестных σ_1^2 и σ_2^2

Очень часто мы не знаем генеральных параметров сравниваемых совокупностей. В этом случае в качестве статистического критерия используют величины, включающие в себя значения выборочных средних, выборочных дисперсий и объемов выборок. Так, например, для проверки статистической гипотезы о равенстве средних арифметических двух нормальных распределений используют t-статистику, которая строится следующим об-

разом. Вначале строится объединенная оценка общей дисперсии σ^2:

$$s_*^2 = \frac{\sum \tilde{X}_1^2 + \sum \tilde{X}_2^2}{n_1 + n_2 - 2},$$

где $\tilde{X}_1 = X_1 - \bar{X}_1$ и $\tilde{X}_2 = X_2 - \bar{X}_2$. Выражение для s_*^2 можно записать иначе:

$$s_*^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{(n_1 - 1) + (n_2 - 1)},$$

где s_1^2 и s_2^2 — выборочные дисперсии по первой и второй выборкам. Предполагая независимость между этими двумя дисперсиями, можно записать дисперсию разности между двумя выборочными средними:

$$s_{\bar{X}_1 - \bar{Y}_2}^2 = \frac{s_*^2}{n_1} + \frac{s_*^2}{n_2} = \frac{n_1 + n_2}{n_1 n_2} \cdot s_*^2 .$$

Выражение для t-статистики можно записать в виде

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{(\bar{X}_1 - \bar{Y}_2)}} = \frac{(\bar{X}_1 - \bar{X}_2) \cdot \sqrt{n_1 n_2}}{s_* \sqrt{n_1 + n_2}}, \qquad (4.3.6)$$

которая в предположении справедливости нулевой гипотезы о равенстве средних подчиняется t-распределению Стьюдента с числом степеней свободы $f = n_1 + n_2 - 2$. При этом для имеющихся выборок значения выборочных дисперсий s_1^2 и s_2^2 не должны статистически различаться при заданном уровне значимости α. Напомним, что это означает следующее:

$$F = (s_1^2/s_2^2) < F_{кр}(\alpha;\ f_1;\ f_2),$$

где $f_1 = n_1 - 1$; $f_2 = n_2 - 1$ и $s_1^2 > s_2^2$. В противном случае сравнение средних надежнее проводить, используя программы однофакторного дисперсионного анализа.

В зависимости от вида альтернативной гипотезы, как и ранее, строятся односторонняя или двухсторонняя критические области. При этом соответственно критические значения t-статистики находятся по таблицам квантилей Стьюдента и равны $t_{\alpha;\ f}$ и $t_{\alpha/2;\ f}$.

В том случае, когда имеются две несгруппированные и необработанные выборки из двух совокупностей и не-

обходимо произвести проверку статистической гипотезы о величине разности между средними этих двух совокупностей, можно использовать программу П 14/34. Величина предполагаемой разности D может принимать различные значения, включая 0, т. е. при $D=0$ мы проверим гипотезу о равенстве средних [7]. Программа позволяет обрабатывать две выборки в случае, если для каждой из них известны дисперсии σ_1^2 и σ_2^2, а также и в том случае, если сами дисперсии неизвестны, но равны. И в этом, и в другом случае искомые t-статистики имеют распределение Стьюдента с $f=n_1+n_2-2$ степенями свободы и позволяют проверить нулевую гипотезу $H_0: \bar{X}_1-\bar{X}_2=D$, где D — предполагаемая разность.

В качестве примера к программе П 14/34 рассмотрим задачу, опубликованную в 1903 году английским физиком Рэлеем.

Пример [39, с. 355]. Рассмотрим данные, полученные при измерении веса определенного объема азота при 15 °С и давлении 760 мм Hg. В первой группе опытов производилось взвешивание азота, полученного из азотистых соединений, во второй — азота, полученного из воздуха (табл. 13).

Таблица 13

Группа опытов			
1-я	1-я	2-я	2-я
2,301 43	2,299 40	2,310 17	2,310 10
2,298 90	2,298 49	2,309 86	2,310 23
2,298 16	2,298 89	2,310 10	2,310 35
2,301 82		2,310 10	2,310 26
2,298 69		2,310 24	2,310 24

Средний вес азота по первой группе опытов равен 2,29947 г, а по второй — 2,31016 г. Таким образом, разность между средними равна 0,01069, т. е. средний вес азота, полученного из воздуха, почти на 1/200 тяжелее азота, полученного из соединений.

Используя программу П 14/34, провести проверку статистической гипотезы о равенстве средних весов азота.

Решение. Формулируем нулевую гипотезу: $H_0: \bar{X}_1 = \bar{X}_2$ при альтернативной $H_1: \bar{X}_1 \neq \bar{X}_2$. После набора программы П 14/34 производим очистку регистров В/О С/П (0) и начинаем ввод данных первой группы измерений: 2,30143 С/П (2,30143) 2,2989 С/П (2,2989) и т. д. По окончании ввода данных из первой группы производим очистку регистров и вводим данные второй группы: В/О С/П (0) 2,31017 С/П (2,31017) и т. д. По окончании ввода второй группы выполняем по п. 6 инструкции: 1 ПО П1. Поскольку нулевая гипотеза H_0 имеет вид $H_0: \bar{X}_1 = \bar{X}_2$, то принимаем $D=0$ в регистр 8 : 0 П8 и вычисляем значение t-критерия: БП 24 С/П (—25,019542=t).

Зададимся значением $\alpha = 5\%$; число степеней свободы для t-критерия $f = n_1 + n_2 - 2 = 8 + 10 - 2 = 16$. По таблице [17, 18] находим значение t для двухстороннего критерия $t_{0,025;\,16} = 2,12$. Поскольку $|t| > t_{0,025;\,16} = 2,12$, то нулевая гипотеза о равенстве средних весов отклоняется. Действительно, вскоре после опытов по взвешиванию азота было установлено, что в азоте, полученном из воздуха, содержится также другой инертный газ, более тяжелый, чем азот, — газ аргон.

4.3.3. Сравнение средних парных наблюдений

Как уже упоминалось раньше, используя результаты наблюдений, т. е. экспериментальные данные, нельзя абсолютно точно доказать нулевую гипотезу. В результате статистической проверки мы можем только установить возможную вероятность того, что мы ошиблись, приняв нулевую гипотезу. Естественно при этом желание исследователя организовать наблюдения и анализ результатов таким образом, чтобы вероятность ошибки минимизировать. Применение теоремы Неймана—Пирсона [2] позволяет производить построение критических областей для используемых критериев таким образом, что ошибка второго рода минимизируется. Вместе с тем и сама организация взятия выборок из генеральных совокупностей также позволяет повысить точность и надежность выводов. До этого мы рассмотрели методы и реализующие их вычислительные программы для сравнения двух независимых выборок. Иногда такие выборки называют также не связанными между собой. Такие выборки представляют собой наборы числовых

данных, отражающие результаты измерений какой-то величины и получены они либо в разное время, либо в одно и тоже время, но на разных рабочих местах и т. п. Иначе говоря, общее в таких выборках только то, что это результаты наблюдения (измерения) за одной и той же величиной.

Часто же наблюдения можно организовать так, что получаемые при этом выборки не являются независимыми. Такие зависимые, связанные между собой выборки представляют собой парные наблюдения, которыми могут быть одновременные наблюдения за температурой печи с помощью двух измерительных устройств, измерения тока в цепи двумя последовательно включенными амперметрами и т. д. В результате парных наблюдений получаются два ряда одинакового объема:

$$X : X_1, X_2, X_3, ..., X_i, ..., X_n;$$

$$Y : Y_1, Y_2, Y_3, ..., Y_i, ..., Y_n.$$

Таким образом, мы имеем n связанных пар (X_i; Y_i). Если имеющиеся различия в организации наблюдений первого и второго ряда (разное оборудование, методики и т. п.) влияют только на один из признаков, X или Y, то между спаренными наблюдениями должно проявиться различие. Тогда задача статистической проверки гипотезы о равенстве заключается в том, чтобы отличить случайные флуктуации между этими двумя величинами от значимо проявляющихся и обусловленных именно имеющимися различиями. Для этой цели находят разности: $d_i = X_i - Y_i$, из которых можно получить среднюю разность.

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i = \frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i), \qquad (4.3.7)$$

где n — число пар наблюдений.

Естественно, что чем меньше $\bar{d}$, тем более вероятна гипотеза о том, что расхождения незначимы и случайны. Следовательно, нулевую гипотезу в таком случае можно сформулировать так:

$$H_0 : \bar{d} = 0. \qquad (4.3.8)$$

В качестве ей альтернативных гипотез может быть выдвинута $H_1 : \bar{d} > 0$, или $H_1 : \bar{d} < 0$ (односторонний крите-

рий), или $H_1: \bar{d} \neq 0$ (двухсторонний критерий). В качестве критерия в этом случае служит t-статистика

$$t = \frac{\bar{d}}{s_{(\bar{d})}}, \qquad (4.3.9)$$

которая подчиняется t-распределению Стьюдента при нормальном распределении d_i с числом степеней свободы $f = n - 1$.

Метод парных сравнений представляет собой более точный и надежный метод сравнения средних двух совокупностей, чем сравнение средних независимых выборок. Это обусловлено тем, что при организации парных наблюдений часть случайных причин, увеличивавших ранее дисперсии наблюдений, не влияет на измеряемую величину. Это приводит к тому, что дисперсия разностей $s^2_{(di)}$ становится меньше аналогичной дисперсии для разности средних $s^2_{(\bar{X}-\bar{Y})}$. Такое уменьшение дисперсии означает повышение точности наблюдений и при прочих равных условиях увеличивает значение t-статистики (4.3.9), что означает увеличение дискриминирующей способности метода. При использовании двух разных методов организации наблюдений может оказаться, что применение метода сравнения двух средних между независимыми выборками дает основание принять нулевую гипотезу, тогда как по методу парных сравнений нулевую гипотезу необходимо отклонить и принять гипотезу о наличии разницы между двумя средними.

Приводимая ниже программа П 16/34 позволяет произвести обработку таких парных наблюдений и вычислить t-статистику [7].

Пример. На основе рационализаторского предложения в цехе собран датчик расхода газов ДРГ. Для оценки его пригодности к эксплуатации необходимо произвести сравнение показаний нового датчика с показаниями серийно выпускаемых датчиков ДРГ-1. Для этой цели оба датчика подключают последовательно к одной газовой магистрали и в один и тот же момент времени снимают показания с обоих датчиков. Число таких измерений — 21. Требуется провести проверку статистической гипотезы о равенстве средних значений показаний обоих датчиков расхода газа с доверительной вероятностью $P_{дов} = 95\%$. Попарные наблюдения за расходом газа в л/ч приведены в табл. 14.

Таблица 14

ДРГ	21,5	20,9	21,4	20,8	21,4	20,1
ДРГ-1	21,2	21,3	21,0	21,5	20,7	20,7
ДРГ	20,8	20,9	21,5	21,2	20,9	20,8
ДРГ-1	20,1	20,4	21,8	21,1	21,5	20,9
ДРГ	21,2	20,8	21,3	21,5	20,8	
ДРГ-1	20,9	20,9	21,7	21,9	21,4	
ДРГ	21,3	21,3	20,4	20,7		
ДРГ-1	21,9	21,1	20,9	20,5		

Решение. Согласно инструкции к П 16/34 вводим попарно все наблюдения и рассчитываем значение t-статистики: $t = -8{,}9856866 \cdot 10^{-1}$. По таблицам находим значение, критическое при $\alpha = 5\%$ (односторонняя гипотеза) $t_{0{,}06;\,20} = 1{,}73$. Поскольку $|t| < t_{0{,}05;\,20} = 1{,}73$, то нулевая гипотеза о равенстве средних принимается. В том случае, когда обе выборки независимых наблюдений обработаны ранее по другим программам и определены $\bar{X}$, s и n, то для сравнения средних можно использовать более короткую программу П 17/34. В этой программе производится вычисление t-критерия Стьюдента согласно (4.3.6).

4.4. Статистическая проверка гипотез о законах распределения

Все рассмотренные выше методы сравнения и использовавшиеся при этом критерии были основаны на предположении о том, что исследуемые случайные величины распределены по некоторому известному закону распределения, чаще всего по нормальному. Как уже говорилось выше, такие методы называются параметрическими.

Исторически сложилось так, что первой была наиболее детально разработана гауссовская статистика, т. е. статистика, связанная с нормальным распределением. Объясняется этот факт тем, что в основном использовать статистические методы обработки и анализа наблюдений стали такие науки, как геодезия и астрономия, ошибки измерений в которых довольно точно подчинялись нормальному распределению. Постепенно все большее количество научных направлений и дисциплин стали использовать количественные методы анализа явлений. При этом стало выясняться, что многие задачи таких дисциплин, как биология, экономика, демография, психология и др. содержат в себе данные, не подчиняющиеся нормальному распределению. К таким данным нельзя было подобрать достаточно точно и другие известные законы распределения. Это и послужило толчком к интенсивному развитию методов непараметрической статистики, которая стала развиваться в основном в послевоенные годы.

В тех случаях, когда известен закон распределения случайной величины, параметрические методы обладают преимуществом перед непараметрическими. Поэтому уже на самых ранних этапах исследования экспериментальных данных и моделирования исследуемого процесса или явления целесообразно установить закон распределения величин, описывающих качество продукции. Исходя из того, что нормальное распределение характеризуется коэффициентами асимметрии и эксцесса, равными нулю, можно по величине выборочных коэффициентов A и E (программы П5/34, П6/34) приближенно оценить нормальность распределения наблюдаемой величины.

Приводимая ниже программа П 19/34 позволяет определить стандартные отклонения для A и E [7]. Сравнивая полученные значения стандартных отклонений s_A и s_E с выборочными значениями A и E, можно сделать оценку нормальности распределения. При этом руководствуются следующим правилом: если наблюдается одновременное выполнение неравенств $|A| \leqslant 3s_A$ и $|E| \leqslant 5s_E$ [2, 27], то распределение можно считать нормальным. При нарушении одного из этих неравенств оснований для этого не имеется.

Выбор коэффициентов 3 и 5 в этих неравенствах связан с применением неравенства Чебышева для оценок доверительных интервалов этих параметров с доверительной вероятностью $P_{дов}=95\%$. Подробный вывод этих коэффициентов дан в [2].

Пример. При обработке выборки объемом $n=105$ были получены выборочные оценки коэффициентов асимметрии и эксцесса $A=-0{,}5$, $E=1{,}3$. Оценить нормальность распределения совокупности, сравнив полученные оценки коэффициентов с их стандартными отклонениями.

Решение. Выполнив инструкцию к программе П 19/34, получаем значение $s_A=2{,}3346806\cdot10^{-1}$ и $s_E=4{,}5391798\cdot10^{-1}$. Поскольку одновременно выполняются оба неравенства $|A|\leqslant3s_A$ и $|E|\leqslant5s_E$, то можно принять гипотезу о нормальности закона распределения. Следует, однако, иметь в виду, что рассмотренные неравенства могут иметь место и для других распределений, отличных от нормального. Поэтому применение данного метода целесообразно применять, главным образом, не для проверки нормальности, а для выявления отклонений распределения от нормального, что не одно и то же.

Более точные методы проверки нормального характера распределения по асимметрии и эксцессу описаны в [17, 18].

4.4.1. Основная гипотеза и критерии согласия

В математической статистике гипотезу о нормальности изучаемого распределения называют основной гипотезой. Статистическую проверку этой гипотезы по выборке производят при помощи критериев согласия. Такие критерии позволяют определить вероятность того, что при выполнении предполагаемого закона распределения в рассматриваемой совокупности, наблюдающиеся в выборке отклонения от этого закона, являются случайными, а не свидетельствуют об ошибочности гипотезы. Если такая вероятность велика, то отклонения от предполагаемого закона признаются случайными, а нулевая гипотеза о законе распределения не опровергается.

Необходимо помнить, что проверка основной гипотезы о нормальности распределения сама по себе не есть самоцель, а необходима чаще всего лишь для обоснова-

ния возможности применения тех или иных статистических критериев.

Отметим вновь, что вероятностный характер таких критериев не дает нам возможности получить абсолютно однозначный ответ. Получаемое при этом заключение также носит вероятностный характер, проявляемый в том, что это заключение мы можем принять с определенной доверительной вероятностью, а следовательно, и с некоторой, хотя и значительно меньшей, вероятностью ошибки этого заключения.

В исследовательской практике применяются самые различные критерии согласия, которые оформлены в виде ГОСТа [20]. Наибольшее применение получили критерии согласия Пирсона, Колмогорова, Смирнова и др. Рассмотрим методику применения критерия согласия Пирсона χ^2 (хи-квадрат) при проверке основной гипотезы о нормальном распределении генеральной совокупности. В основе метода сравнения по критерию χ^2 лежит сравнение фактически наблюдаемых частот с теоретическими, которые вычисляются в предположении нормального распределения. Как правило, эти частоты отличаются друг от друга. Например, распределение эмпирических частот, полученное при обработке выборки объемом $n=200$ с помощью программы П1/34, имеет следующий вид:

8; 16; 40; 72; 36; 18; 10.

Соответствующие теоретические частоты вычисляем с применением функции Лапласа $\Phi_0(z)$ [17], где z — нормированная переменная. При чем $z_i=(X_{i+1}-\bar{X})/s$. Наименьшее значение $z_i=z_1$ получают равным $-\infty$, а наибольшее $+\infty$. Тогда вероятности (теоретические) попадания в i-тый интервал вычисляются по следующей формуле:

$$P_i(X_i < X \leqslant X_{i+1}) = \Phi_0\left(\frac{X_{i+1}-\bar{X}}{s}\right) - \Phi_0\left(\frac{X_i-\bar{X}}{s}\right) = \Phi_0(z_{i+1}) - \Phi_0(z_i). \quad (4.4.1)$$

После того как полученные значения теоретических вероятностей попадания умножим на весь объем $n=200$, получим значения теоретических частот:

6; 18; 36; 73; 39; 18; 7.

Как видно, эмпирические и теоретические частоты попадания величины X в семь интервалов отличаются друг от друга. Случайно ли это? Не исключено, что такое расхождение объясняется малым объемом выборки, а возможно и тем, что мы ошибаемся, предполагая нормальность распределения, когда вычисляем теоретические частоты. В таком случае это означает, что нужно использовать другой закон распределения, тот, которому подчиняется генеральная совокупность. Но мы его не знаем. Выход из этого замкнутого, казалось бы, круга и состоит в том, что мы выдвигаем основную гипотезу о виде закона распределения в генеральной совокупности и проверяем ее.

В качестве критерия для проверки основной гипотезы используют случайную величину

$$\chi^2_k = \sum_{j=1}^{k} \frac{(n_i - n_i')^2}{n'_i}. \qquad (4.4.2)$$

Случайной эта величина является вследствие случайности выборки и значений фактических частот n_i. Здесь через n'_i обозначены теоретические частоты попадания в интервал; k — число интервалов. Чем ближе друг к другу фактические и теоретические частоты для каждого интервала, тем меньше величина χ^2_k, а это и служит признаком близости фактического и предполагаемого законов распределения. Следует обратить внимание на тот факт, что независимо от того, какому закону распределения подчинена генеральная совокупность, величина, определяемая выражением (4.4.2), стремится к закону распределения Пирсона χ^2 с l степенями свободы при $n \to \infty$. Это означает, что критерии χ^2 можно применить для проверки не только нормального закона, но и других. Именно вследствие связи с законом распределения Пирсона критерий (4.4.2) обозначают как χ^2_k. Число степеней свободы l определяют по выражению $l = k - 1 - r$, где k — число интервалов; r — число параметров предполагаемого закона распределения, которые оцениваются по выборке. Поскольку в выражение плотности вероятности нормального закона входят 2 параметра — среднее и дисперсия, то для нормального закона $l = k - 3$. Односторонняя критическая область строится из предположения о том, что вероятность попадания критерия χ^2 в эту область (область отклонения), когда верна ос-

новная гипотеза (она же нулевая гипотеза о равенстве двух законов распределения), должна быть равна принятому уровню значимости α:

$$P\{\chi^2_k > \chi^2_{кр}(\alpha, l)\} = \alpha. \qquad (4.4.3)$$

Это означает, что при $\chi^2_k > \chi^2_{кр}(\alpha, l)$ нулевая гипотеза отвергается, а при $\chi^2_k < \chi^2_{кр}$, (α, l) она принимается. Значение $\chi^2_{кр}$ определяют по таблице распределения χ^2.

При использовании критерия χ^2 необходимо, чтобы объем выборки был не менее 50 и в каждом интервале частота была не менее 5—8. Малочисленные группы необходимо объединять в одну, частота в которой не меньше 5—8. Для проверки результатов вычисления статистики χ^2_k можно использовать два тождественных выражения:

$$\chi^2_k = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \qquad (4.4.4)$$

$$\chi^2_k = \sum_{i=1}^{k} \frac{n^2_i}{n^2_i} - n = \left(\frac{1}{n} \sum_{i=1}^{k} \frac{n^2_i}{p_i}\right) - n, \qquad (4.4.5)$$

где p_i — расчетные значения вероятности попадания величины X в i-тый интервал. С использованием (4.4.5) и составлена программа П 20/34 [7].

Пример. При разработке технологии золочения металлической ленты, используемой в производстве высокочастотных транзисторов, методом плакирования требовалось установить закон распределения толщины золотого покрытия в партии. Это требование со стороны заказчика разработки мотивировалось тем, что знание закона распределения позволяет ему более достоверно оценить доверительные интервалы, толерантные интервалы и качество партии ленты в целом. Кроме того, это позволяло внедрить на операции контроля толщины слоя золота ГОСТ 20 736—75 «Качество продукции. Статистический приемочный контроль по количественному признаку при нормальном распределении контролируемого параметра», если закон распределения окажется нормальным.

При проверке гипотезы нормальности закона распределения был использован критерий согласия Пирсона χ^2. В табл. 15 приведены результаты измерений в виде распределения фактических частот n_i, а также расчет-

ных вероятностей P_i в предположении нормальности закона распределения для одной партии.

Требуется проверить основную гипотезу с уровнем значимости $\alpha = 5\%$.

Таблица 15

Номер интервала i	1	2	3	4	5	6	7
Фактическая частота n_i	8	16	40	72	36	18	10
Расчетная вероятность P_i	0,03	0,09	0,18	0,38	0,195	0,09	0,035

Решение. Выполняя инструкцию к программе П 20/34, вводим в ПМК имеющиеся данные. В результате обработки данных получаем значение $\chi^2_7 = 3{,}06034 \approx 3{,}1$. Находим число степеней свободы $l = k - 3 = 7 - 3 = 4$. По таблице [17, 18] находим критическую точку $\chi^2_{кр}(0{,}05;\ 4) = 9{,}49$. Поскольку $\chi^2_{кр}(0{,}05;\ 4) > \chi^2_7$, то основную гипотезу о нормальном законе распределения толщины слоя золотого покрытия на ленте принимаем. Применение программы П 20/34 требует предварительного вычисления теоретических, расчетных значений вероятностей P_i попадания наблюдений в i-тый интервал. Такой расчет может быть с достаточно высокой степенью точности выполнен на ПМК. Объединяя расчет значений P_i для нормального закона распределения и χ^2_k в одну, большую по объему, программу, получим программу П 21/34.

Кроме рассмотренного метода сравнения χ^2_k с $\chi^2_{кр}(\alpha,\ l)$ часто применяют и другой метод проверки основной гипотезы. На основе полученного χ^2_k и l по таблицам определяют вероятность $P(\chi^2_{кр} \geqslant \chi^2_k)$. Фактически это означает нахождение уровня значимости, при котором расчетное значение χ^2_k равно критическому $\chi^2_{кр}(P,\ l)$. Отсюда следует правило проверки: если величина вероятности $P\ (\chi^2_{кр} \geqslant \chi^2_k)$, называемая также значимостью, больше, чем выбранный уровень значимости, то нулевая гипотеза принимается, и наоборот. Рассмотрим несколько подробнее смысл такой проверки, т. к. ниже приводится программа П 22/34, реализующая схему [7]. Предположим, что мы задались уровнем зна-

чимости $\alpha=5\%$, соответственно доверительная вероятность $P_{дов}=100-5=95\%$. Используя приведенные выше значения фактических и расчетных частот для семи интервалов в примере к программе П 20/34, определим значение χ^2_7:

$$\chi^2_7 = \sum_{i=1}^{7} \frac{(n_i - n_i')^2}{n_i'} \approx 3{,}1 . \qquad (4.4.6)$$

Определим число степеней свободы $l=k-3=7-3=4$. По таблицам [50] найдем вероятность $P(\chi^2_{кр} \geqslant \chi^2_7) = 0{,}5578$. Это означает, что вероятность того, что расчетное значение критерия согласия χ^2_7 меньше или равно критическому значению $\chi^2_{кр}$ очень высока. Вспомним, что чем меньше величина $\chi^2_{расч}$, тем ближе фактическое распределение к теоретически предполагаемому. Поскольку вероятность $P(\chi^2_{кр} \geqslant \chi^2_7) = 0{,}5578 > 0{,}05 = \alpha$, то мы принимаем нулевую гипотезу. Напротив, если бы мы получили значение $P(\chi^2_{кр} \geqslant \chi^2_7) = 0{,}01 = 1\%$, то это означало бы, что отклонения в частотах случайны, а не являются проявлением того, что фактический закон распределения в генеральной совокупности иной, нежели тот, по которому производился подсчет теоретических частот. Поскольку эта величина меньше принятой нами 0,05, то нулевую гипотезу следовало бы отклонить.

Для использования программы П 22/34 при проверке нормальности распределения генеральной совокупности по выборке необходимо знать следующие выборочные параметры: среднее $\bar{X}$ и дисперсию s^2, которые можно определить с помощью ряда предыдущих программ. При обработке вводимого массива данных производится вычисление величины χ^2_k и нахождение вероятности $P(\chi^2_{кр} \geqslant \chi^2_k)$ (значимости). В случае $P > \alpha$ нулевая гипотеза принимается, при $P < \alpha$ нулевая гипотеза о нормальности совокупности отвергается.

Значения критических точек $\chi^2_{кр}(\alpha, f)$ и значения вероятностей $P(\chi^2_{кр} \geqslant \chi^2_k)$ приведены полностью в [17].

Для контроля программы П 22/34 можно воспользоваться контрольным примером к программе П3/34: выборки 1; 2; 3 имеет $\bar{X}=2$ и $s^2=1$. Выполнив инструкцию к П 22/34, получаем значение $P(\chi^2_{кр} \geqslant \chi^2) = 4{,}2319001 \cdot 10^{-1}$.

4.4.2. Проверка однородности наблюдений

Поскольку при проверке статистических гипотез наши заключения имеют вероятностный характер, то это предъявляет строгие требования к качеству имеющихся в нашем распоряжении данных, которые должны быть однородны. Гипотеза однородности таких данных предполагает, что они принадлежат одной генеральной совокупности. При невыполнении этого требования никакие хитроумные и сложные методы и формулы в сочетании с самыми совершенными ПМК и ЭВМ не позволят нам получить надежные заключения о свойствах процесса и промоделировать его. «Математическую формулу действительно можно уподобить мельничному жернову, из-под которого нельзя получить пшеничную муку, если положить под него вместо зерна лебеду» [21].

При проверке гипотезы однородности двух выборок широкое применение получили сериальные непараметрические критерии. Суть этих критериев в том, что из двух имеющихся независимых выборок составляют одну выборку, которую располагают в порядке возрастания случайной величины. В этом упорядоченном вариационном ряду выделяют серии расположенных подряд элементов одной и той же выборки. Далее производится подсчет количества l_i элементов в сериях. Ряд полученных чисел l_i и используют для построения непараметрического критерия. В программе П 24/34 использован один из таких сериальных критериев T* [7, 17].

Ранее мы уже упоминали о преимуществах парных наблюдений по сравнению с использованием независимых выборок. Полученные парные наблюдения могут быть обработаны для проверки их однородности с помощью довольно простого и достаточно эффективного критерия знаков Диксона и Муда. Рассмотрим суть метода, а затем его программную реализацию и возможные применения.

Предположим, что проведено n пар измерений некоторой случайной величины при отличающихся условиях. Одно из значений пары получено при условиях A, а другое при условиях B. Разности этих пар обозначим $\Delta_i = A_i - B_i$. Значения Δ_i могут быть отрицательными, равными нулю, и положительными. Как правило, нулевых значений Δ_i наблюдается мало; если их доля сос-

тавляет единицы процентов от общего объема пар, то их можно опустить из дальнейшего рассмотрения либо поделить поровну между положительными и отрицательными значениями Δ_i. Очевидно, что если законы распределения для обеих выборок совпадают, то разности Δ_i в среднем не должны отличаться от нуля, т. е. около половины $\Delta_i>0$, а другая половина $\Delta_i<0$. В этом случае вероятность появления определенного числа $\Delta_i>0$ или, наоборот, $\Delta_i<0$ для данного общего объема пар наблюдений подчиняется биноминальному распределению с $p=q=0{,}5$. Поскольку используемый критерий знаков основан на биноминальном распределении, то результаты отдельных пар наблюдений должны быть независимыми друг от друга, т. е. знак разности для одной пары измерений никак не должен влиять на знак разности любой другой пары. Критерий знаков проверяет нулевую гипотезу о том, что медианы распределения разностей равны нулю. Для того чтобы определить искомую вероятность появления m или более значений $\Delta_i>0$ из общего числа пар n, используют выражение

$$P_n(m) = 2^{-n} \sum_{i=m}^{n} C_n^{\,i}, \qquad (4.4.7)$$

где C^i_n — число сочетаний из n элементов по i.

$$C_n^{\,i} = \frac{n!}{i!\,(n-i)!}. \qquad (4.4.8)$$

Для того чтобы результат проверки нулевой гипотезы при двухстороннем критерии был надежен на 5%-ном уровне, необходимо минимум 6 пар наблюдений. Эффективность критерия знаков уменьшается с ростом объема выборки от 0,95 при $n=6$ до 0,64 при $n\to\infty$ по сравнению с параметрическим t-критерием Стьюдента при условии нормального распределения разностей Δ_i.

Рассмотрим более подробно формулировку альтернативной гипотезы H_1 и условия принятия и отклонения нулевой гипотезы. Условимся обозначать через m число случаев чаще встречающегося знака. Если это число положительных разностей $\Delta_i>0$, то в этом случае логичнее выдвигать альтернативную гипотезу H_1^+ в том, что среднее значение совокупности A больше среднего значения совокупности B. Если же m — это число отрицательных разностей $\Delta_i<0$, то соответственно логичнее

выдвигать альтернативную гипотезу H_1^- о том, что среднее значение совокупности А меньше среднего значения совокупности В. При $m > \frac{n}{2}$ значение $P_n(m) < 50\%$. Поэтому при использовании одностороннего критерия, т. е. при альтернативной гипотезе H_1^+ или H_1^-, нулевая гипотеза принимается, если $P_n(m) > \alpha$, и наоборот, H_0 отвергается, если $P_n(m) \leqslant \alpha$. Если же о знаке возможной разности средних значений совокупностей А и В не выдвигается никакого предположения, т. е. альтернативная гипотеза формулируется в виде $H_1 : \bar{X}_A - \bar{X}_B \neq 0$, то используется двухсторонний критерий знаков. В этом случае нулевая гипотеза принимается в том случае, если $P_n(m) > \frac{\alpha}{2}$.

Рассмотренный метод непараметрического сравнения с помощью критерия знаков Диксона и Муда реализован в программе П 25/34 [7].

Пример. Проводилось изучение влияния предпосевной обработки семян зернового злака на урожайность. Для проверки гипотезы о том, что обработка семян повышает урожайность, было засеяно 44 опытных делянки и 44 контрольных делянки. С целью устранения влияния других факторов опытная и контрольная делянки располагались рядом и имели равные площади. Вся обработка до уборки урожая также была одинакова на тех и других делянках. После уборки урожая было установлено, что в 27 парах смежных делянок имелась положительная разность урожая $\Delta_i = O_i - K_i > 0$, т. е. на опытных делянках урожай был выше, чем на соседних контрольных. На остальных 14 парах делянок эта разность была отрицательна.

Требуется проверить альтернативную гипотезу H_1^+ о том, что средний урожай на опытных делянках с использованием обработанных семян выше, чем на контрольных. Уровень значимости для проверки этой односторонней гипотезы взять равным 5%.

Решение. Выполняя инструкцию к программе П 25/34, после ввода программы в ПМК и перевода ее в автоматический режим выполняем: 27↑44 В/О С/П. Продолжительность вычисления P_{44} (27) составляет около 73 секунд. Результат $P_{44}(27) = 8{,}7085683 \cdot 10^{-2} \approx 9\%$. Посколь-

ку $P_{44}(27) > 5\%$, то альтернативная гипотеза H_1^+ отвергается и принимается гипотеза H_0, т. е. оснований утверждать о том, что использовавшаяся предпосевная обработка семян приводит к увеличению среднего урожая, нет.

4.5. Проверка статистических гипотез методом последовательного анализа

При решении примеров к рассмотренным выше методам нередко оказывалось, что величина статистического критерия значительно больше, либо значительно меньше соответствующей заданному уровню значимости критической величины. Вернемся к последнему нашему примеру, к программе П 25/34. Допустим, что последняя разность $\Delta_{44} < 0$, т. е. $O_{44} < K_{44}$. Тогда имеем, что после проведения 43 пар измерений было 27 положительных разностей и 16 отрицательных. Посчитаем вероятность этого события P_{43} (27) и сравним ее с $\alpha = 5\%$. Оказывается, что $P_{43}(27) = 6{,}3144825 \cdot 10^{-2} \approx 6\% > \alpha$. Если же предположить, что последняя разность Δ_{44} была положительной, тогда посчитаем P_{43} (26) и также сравним ее с α. В этом случае имеем $P_{43}(26) = 1{,}1102656 \cdot 10^{-1} \approx \approx 11\% > \alpha$. Выходит, что и в том, и в другом случае уже после 43 пар измерений имелось достаточно информации, чтобы, используя исходные данные, принять нулевую гипотезу. Смысл этих рассуждений сводится к тому, что при решении рассмотренных выше задач сравнения и сопоставления мы нередко имеем избыточную информацию по сравнению с минимально необходимой для решения конкретной задачи. При использовании методов с фиксированным числом экспериментов (наблюдений, измерений) мы производим только один раз обработку результатов и получаем один из двух ответов на поставленные гипотезы: нулевая гипотеза принимается или нулевая гипотеза отвергается.

Во время второй мировой войны американский статистик Абрахам Вальд разработал новый, последовательный анализ, предназначенный для проверки статистических гипотез. Впервые идея такого вида проверки статистических гипотез была высказана другими учеными— Доджем и Ромигом в 1929 г. В связи с проводимыми конструкторскими разработками в области военно-мор-

ского оборудования А. Вальд развил эту идею в цельную и стройную теорию последовательного анализа. Первый отчет по этим работам был сделан А. Вальдом в сентябре 1943 г., второй — в июле 1944 г., а в 1947 г. в США была опубликована его книга «Последовательный анализ», переведенная и изданная в СССР в 1960 г. При использовании этого нового метода измерения проводились в виде последовательной серии. После каждого испытания производился анализ всех имеющихся результатов измерений, на основании которого делалось одно из трех следующих заключений:

1. Проверяемая гипотеза принимается.

2. Проверяемая гипотеза отвергается.

3. Имеющейся информации для принятия первого или второго заключения недостаточно, измерения необходимо продолжать.

Вплоть до окончания второй мировой войны последовательный анализ оставался военной тайной, так как он тотчас был признан наиболее рациональным методом непрерывного контроля качества в промышленном производстве. Последовательный анализ позволяет сокращать число необходимых экспериментов в среднем в 2 раза по сравнению с обычными методами, в которых число наблюдений не является случайной величиной, как в последовательном анализе, а заранее определенным, фиксированным числом.

Программа П 26/34 реализует последовательный анализ для проверки статистической гипотезы о равенстве средних, программа П 27/34 — последовательную проверку гипотез о дисперсии с известным значением генерального среднего, а программа П 28/34 — с неизвестным значением генерального среднего.

Пример. Для проверки установки измерения электрической емкости конденсаторов на заводе используется эталонный образец емкостью 1 мкФ. При данной емкости согласно паспорту на установку $\sigma_0=0{,}1$ мкФ. При проверке электрической емкости в конденсаторах одной из партий возникли сомнения в точности измерений на данной установке. Требуется провести проверку фактической дисперсии с использованием эталона, приняв $\sigma_1=0{,}3$ мкФ, $\alpha=5\%$, $\beta=1\%$.

Ниже приведены результаты 13 измерений эталонного образца (мкФ): 0,9; 1,1; 1,2; 0,95; 1,3; 0,8; 0,9; 1,0; 1,0; 1,4; 1,1; 1,2; 0,8.

Решение. Принимая $\mu_0=1$ мкФ, $\sigma_0=0{,}1$ мкФ и $\sigma_1=0{,}3$ мкФ, формулируем две конкурирующие статистические гипотезы: $H_0: \sigma^2 \leqslant \sigma_0^2$ и $H_1: \sigma^2 \geqslant \sigma_1^2$. После ввода программы П 24/34 вводим исходные данные задачи: 0,1 П7 0,3 П8 1 П9 0,05 П5 0,01 П6. Выполняем п. 3 инструкции: В/О С/П ($6{,}7177841 \cdot 10^{-2}$). Далее начинаем ввод результатов измерения емкости эталонного образца на установке: 0,9 С/П (0) 1,1 С/П (0) 1,2 С/П (0) 0,95 С/П (0) 1,3 С/П (0) 0,8 С/П (0) 0,9 С/П (0) 1,0 С/П (0) 1 С/П (0) 1,4 С/П (—1). Последний результат анализа —1 позволяет принять гипотезу о том, что $\sigma=0{,}3$ мкФ, т. е., действительно, точность измерения на данной установке не соответствует паспортному значению $\sigma_0=0{,}1$.

Для контроля программы П 28/34 можно использовать пример к предыдущей программе П 27/34. Результаты анализа идентичные, после анализа десятого значения емкости 1,4 индицируется —1, т. е. H_0 отвергаем.

ЗАДАЧИ К ГЛ. 4

1. Проверялась работа токарного станка-автомата. По выборке из 13 втулок была оценена дисперсия диаметра втулок $s_1^2=10{,}24$ мкм^2. Было принято решение провести профилактику станка и его наладку. После окончания профилактики и наладки была взята выборка из 18 втулок и определена дисперсия $s_2^2=4{,}84$ мкм^2. Требуется определить, привела ли профилактика и наладка станка к увеличению точности изготовления деталей. Значение $\alpha=5\%$.

2 [40, с. 196]. Для проверки эффективности новой технологии отбираются две группы рабочих: в первой группе, численностью $n_1=40$ чел., где применяется новая технология, получены следующие данные: средняя выработка в штуках $\bar{X}_1=84$ при $s_1=10{,}1$, во второй группе, численностью $n_2=54$ чел., $\bar{X}_2=77{,}5$; $s_2=8{,}4$. Определить с уровнем значимости $\alpha=0{,}05$, действительно ли новая технология оказала влияние на среднюю производительность.

3 [25, с. 214]. В результате анализа газа из двух различных источников были получены следующие данные о содержании метана (мольный процент):

Источник 1: 64,0; 65,0; 75,0; 67,0; 64,5; 74,0; 75,0.

Источник 2: 69,0; 69,0; 61,5; 67,5; 64,0.

Установить, является ли значимым различие в содержании метана в этих источниках. Уровень значимости принять равным 5%.

4 [25, с. 146]. Два разных сорта бензина использовались для определения числа пройденных километров по шоссе на литр израсходованного бензина. Каждый сорт бензина (октановые числа 90 и 94) использовался в пяти идентичных автопробегах по одному и тому же маршруту; были получены следующие результаты:

	94	90
	октановые числа	
Выборочное среднее, км/л	22,7	21,3
Выборочный стандарт, км/л	0,45	0,55

Различны ли эти сорта при уровне значимости $\alpha = 0{,}05$? Если так, то является ли бензин с октановым числом 94 значимо лучшим по сравнению с бензином с октановым числом 90?

5. Из партии конденсаторов, выпущенных в цехе за одну смену, взята выборка в количестве 199 шт. Требуется проверить, подчиняется ли распределение электрической емкости выпускаемых конденсаторов нормальному закону. Результаты измерений сгруппированы в 10 интервалов (табл. 16).

Здесь i — номер интервала; n_i — число наблюдений в i-том интервале; P_i — расчетная вероятность попадания наблюдений в i-тый интервал, вычисленная в предположении нормального закона распределения.

6 [43, с. 155]. В табл. 17 приведены данные опыта по изучению влияния определенной предпосевной обработки семян пшеницы на урожайность. В соответствии с погодными условиями урожайность меняется год от года как в опыте, так и в контроле; поэтому значения опытной группы и значения контрольной группы нельзя

Таблица 16

i	1	2	3	4	5	6	7	8	9	10
n_i	7	11	16	27	43	39	30	14	7	5
P_i	0,0235	0,0475	0,0975	0,158	0,2015	0,1945	0,146	0,078	0,0385	0,015

Таблица 17

Год	1947	1948	1949	1950	1951	1952	1953
Опыт	22,9	20,2	19,5	30,5	35,6	31,9	27,7
Контроль	19,4	16,2	16,9	29,3	31,4	28,5	25,6

Таблица 18

X_1, °C	36,8	37,1	37,3	37,0	37,1	36,9	36,7	37,2	37,0	36,9
X_2, °C	36,9	37,2	37,2	37,2	37,3	37,0	36,8	37,2	37,2	37,1

Таблица 19

X_1	3,8	4,2	5,0	4,8	3,6	4,2	4,6	4,0	4,8	5,2	4,0	4,2	4,8	5,2	5,0	4,6
X_2	3,6	4,4	5,2	5,2	4,0	4,3	5,0	4,4	4,9	5,3	4,2	4,1	5,0	5,5	5,1	4,8

считать взаимно независимыми — они попарно связаны тем, что значения из каждой пары относятся к одному и тому же году и, следовательно, к одним и тем же условиям погоды (которые, как это видно из табл. 17, оказывают на урожайность гораздо большее влияние, чем предпосевная обработка семян).

Требуется провести проверку гипотезы о равенстве средних урожаев в опытной и контрольной группах при $P_{дов}=90\%$.

7 [50, с. 146]. Измерены температура тела у 10 новорожденных детей под мышкой (X_1) и в прямой кишке (X_2). Получены следующие результаты (табл. 18).

Проверить нулевую гипотезу $H_0: \mu_1=\mu_2$ при $\alpha=5\%$.

8 [50, с. 151]. Подлежит проверить точность, с которой работают два лаборанта. Им поручается определить число эритроцитов в крови, взятой у одних и тех же лиц (что им не известно). Каждый из них производит по 21 исследованию. На полученных результатах было вычислено среднее квадратическое отклонение. У первого лаборанта этот показатель равен $s_1=100000$, а у второго — $s_2=60000$. Можно ли предположить, что второй лаборант работает более точно, чем первый? Значение $\alpha=5\%$.

9 [50, с. 75]. В Институте акушерства и гинекологии при изучении гипогалактии (уменьшение выделения молока у кормящих матерей при неправильном питании, инфекциях, интоксикациях, мастите и др.) была дана задача проверить, насколько результаты отдельных составных веществ материнского молока, полученные в лаборатории института, отличаются существенно от тех же результатов, полученных в Институте питания. Были получены следующие параллельные результаты, касающиеся содержания жиров в материнском молоке (табл. 19).

Здесь X_1 — результаты анализов, сделанных в Институте акушерства и гинекологии, X_2 — результаты анализа Института питания (%).

Проверить с помощью программы П 16/34 гипотезу о равенстве средних при $\alpha=5\%$.

РЕШЕНИЯ.

1. Формулируем нулевую гипотезу $H_0: \sigma^2_1 = \sigma_2^2$. Поскольку увеличение точности означает, что одна дисперсия больше другой, то альтернативную гипотезу формулируем следующим образом: $H_1: \sigma_1^2 > \sigma_2^2$, т. е. гипотеза односторонняя. Находим F-отношение: $F = s_1^2/s_2^2 = 2{,}11$. Табличное значение $F_{0{,}05;\ 12;\ 17} = 2{,}4$. Поскольку табличное значение больше F-отношения, то у нас нет оснований отклонить нулевую гипотезу и мы ее принимаем, т. е. увеличения точности не произошло.

2. После набора программы П 17/34 вводим исходные данные: 84 ПО 77,5 П1 40 П2 54 П3 10,1 П4 8,4 П5. Вычисляем значение t-критерия: В/О С/П ($3{,}4018555 = t_{расч}$). Критическое табличное значение t-критерия при $\alpha = 0{,}05$ и $f = 92$ равно 1,98. Так как $t_{расч} > 1{,}98$, то делаем вывод о том, что у нас нет оснований принять нулевую гипотезу $\mu_1 = \mu_2$. Это означает, что новая технология действительно увеличивает среднюю производительность труда.

3. Используя программу П3/34, находим s_1 и s_2 для источника 1 и источника 2 : $s_1 = 5{,}1950135$, $s_2 = 3{,}3279122$. Для проверки гипотезы о равенстве дисперсий σ_1^2 и σ_2^2 находим отношение s_1/s_2 и возводим его в квадрат, получаем значение $F = s_1^2/s_2^2 = 2{,}4368546 \approx 2{,}437$. По таблице [17, 18] находим $F_{0{,}05;\ 6;\ 4} = 6{,}2$. Поскольку $F_{0{,}05;\ 6;\ 4} > F = 2{,}437$, то принимаем нулевую гипотезу о равенстве дисперсий по источнику 1 и 2. Это позволяет использовать программу П 14/34. Выполнив инструкцию к этой программе, получим, что $t = 1{,}1335817 \approx 1{,}13$. По таблице [17, 18] находим $t_{0{,}05;\ 10} = 2{,}23$. Поскольку $t_{0{,}05;\ 10} > t = 1{,}13$, то нулевая гипотеза о равенстве содержания метана в источниках 1 и 2 принимается.

4. Для проверки первого предположения формулируем нулевую и альтернативную гипотезы: $H_0: \mu_{94} = \mu_{90}$; $H_1: \mu_{94} \neq \mu_{90}$. Таким образом, альтернативная гипотеза является двухсторонней. Используя программу П 17/34, находим значение t-критерия: $t_{расч} = 4{,}4052172 \approx 4{,}41$. По таблице [17, 18] находим табличное $t_{0{,}025;\ 8} = 2{,}31$. Поскольку $t_{расч} > 2{,}31$, то нулевая гипотеза отвергается, т. е. эти сорта действительно значимо различаются.

Для проверки второго предположения формулируем другую альтернативную гипотезу: $H_1 : \mu_{94} > \mu_{90}$. По таблице находим $t_{0,05;\,8} = 1,86$. Поскольку $t_{расч} > 1,86$, то нулевую гипотезу отвергаем и принимаем альтернативную, т. е. бензин с октановым числом 94 действительно значимо лучше бензина с октановым числом 90.

5. Используя программу П 20/34, вычислим значение статистики χ^2_{10}. После ввода программы производим очистку регистров В/О С/П (0) и начинаем ввод таблицы: 0,0235↑7 С/П (2085, 1063) и т. д. 0,015↑5 С/П (40484,869). Ввод таблицы окончен. Вычислим значение статистики χ^2_{10}: БП 14 С/П ($4,44155 = \chi^2_{10}$). Число степеней свободы $f = 10 - 3 = 7$. По таблице [17, 18] находим $\chi^2_{0,05;\,7} = 14,03$. Так как найденное по выборке значение статистики χ^2_{10} меньше соответствующего табличного, то принимаем гипотезу о нормальном законе распределения электрической емкости конденсаторов.

6. После набора программы П 16/34 и очистки регистров по п. 2 инструкции к П 16/34 начинаем ввод данных таблицы: 22,9↑19,4 С/П (12,25 — результат после останова) 20,2↑16,2 С/П (28,25) 19,5↑16,9 С/П (35,01) 30,5↑29,3 С/П (36,45) 35,6↑31,4 С/П (54,09) 31,9↑28,5 С/П (65,65) 27,7↑25,6 (70,06) БП 16 С/П (7,3171769), т. е. $t \approx 7,32$. Число степеней свободы $f = n - 1 = 7 - 1 = 6$. По таблице [17, 18] находим значение $t_{0,1;\,6} = 3,71$. Поскольку $t_{расч} > t_{0,1;\,6}$, то различие следует считать значимым.

7. Результат обработки данных по П 16/34 : $t = -3,497993$. Можно ли принять нулевую гипотезу с $\alpha = 5\%$?

8. Формулируем нулевую и альтернативные гипотезы:

$$H_0 : \sigma_1^2 = \sigma^2_2,$$
$$H_1 : \sigma_1^2 > \sigma_2^2.$$

Находим F-отношение:

$$F = \left[\frac{100000}{60000}\right]^2 = 2,78.$$

9. Используя программу П 16/34, получаем значение t-критерия: $t = -4,2944875$. Сравните это значение с табличным и сделайте вывод относительно нулевой и альтернативной гипотез.

ГЛАВА 5

АНАЛИЗ СТАТИСТИЧЕСКОЙ СВЯЗИ МЕЖДУ КАЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

Начала истины в естествознании
должны утверждаться наблюдениями.

К. Линней

5.1. Основные типы шкал измерения

Нередко исследуемые свойства тех или иных объектов не представляется возможным измерить в какой-либо непрерывной шкале, например в шкале отношений. Это бывает в тех случаях, когда не существует шкалы измерения, в которой бы можно было количественно измерить степень проявления исследуемого свойства у данного объекта исследования. В этих случаях принято говорить о качественных оценках этих свойств. Такая же ситуация возникает и тогда, когда данное свойство имеет строго определенное количество отличающихся друг от друга значений. Для описания таких свойств наиболее широко в практике используют две шкалы измерений: номинальную и порядковую. Рассмотрим каждую из этих шкал.

Номинальная шкала используется тогда, когда требуется отнести объект исследования по исследуемому свойству, признаку в один из заранее определенных классов. Примером такого свойства может быть пол животного. В этом случае объекту исследования — животному, приписывается одно из двух значений: буква М или Ж, число 1 или 2, либо какие-то другие условные метки. При наличии только двух классов, как, например, в случае с полом животного, принято говорить о дихотомических переменных.

Следует отметить, что арифметические операции над такими величинами не имеют смысла. Именно по этой причине среднее арифметическое и медиана не могут выступать в данном случае в качестве меры положения центра. В данной шкале наиболее разумно использование для этой цели моды.

По сравнению с номинальной шкалой в технических приложениях большее распространение получила порядковая шкала. Отличие шкалы порядка от номинальной заключается в том, что классы, в которые необходимо относить объект по наличию в нем исследуемого свойства, упорядочены, проранжированы по степени выраженности исследуемого свойства. Например, в номинальной шкале бессмысленно говорить о том, что в женской особи признак пола выражен сильнее, нежели в мужской особи, либо наоборот. А в порядковой шкале, например в шкале твердости Мооса, можно говорить о том, что твердость алмаза (10 единиц) выше твердости талька (1 единица) или гипса (2 единицы).

В данном случае шкала Мооса имеет 10 упорядоченных классов, рангов. Обычно принято нумеровать используемые классы. Еще одним примером порядковой, ранговой шкалы является пятибалльная шкала оценки успеваемости учащихся. Выбор числа классов и их оцифровку производит сам исследователь. Нередко порядковую шкалу используют в том случае, когда показатель свойства исследуемого объекта отражает большое количество более детальных признаков этого объекта и является его интегральным показателем. В этом случае нередко для такого показателя используют четыре градации: «плохо», «удовлетворительно», «хорошо», «очень хорошо» и производят следующую оцифровку этих групп: 1, 2, 3, 4 или наоборот, 4, 3, 2, 1. Следует отметить, что оцифровка качественных переменных (дуальное шкалирование) является достаточно сложной статистической задачей с точки зрения ее оптимальности при поиске зависимостей между порядковыми переменными [3, 38].

5.2. Ранговая корреляция Спирмэна

В том случае, когда изучаемый объект имеет два или более свойств, которые измеряются в порядковой шкале, может возникнуть задача о выявлении статистической связи между этими порядковыми переменными. В этом случае используется понятие ранговой корреля-

ции, величина которой выражается коэффициентом ранговой корреляции Спирмэна r_s [3, 19]:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \qquad (5.2.1)$$

где d_i — разность значений рангов, расположенных в двух рядах у одного и того же объекта. Коэффициент ранговой корреляции r_s является частным случаем коэффициента парной корреляции r_{XY}. Аналогично значения r_s находятся внутри интервала $1 \geqslant r_s \geqslant -1$. С увеличением объема выборки r_s стремится к генеральному параметру ρ_s, но разность между ними не превышает 0,018. Нулевую гипотезу $\rho_s = 0$ проверяют с помощью t-статистики (при $n \geqslant 10$):

$$t = \frac{r_s \sqrt{n-2}}{\sqrt{1 - r_s^2}}, \qquad (5.2.2)$$

имеющей распределение Стьюдента с $f = n - 2$ степенями свободы. При альтернативной гипотезе $\rho_s \neq 0$ используют двухсторонний критерий, т. е. значение t по (5.2.2) сравнивают с табличным $t_{\alpha/2;\,(n-2)}$, а при альтернативных гипотезах $\rho_s > 0$ или $\rho_s < 0$ используют односторонний критерий, т. е. сравнивают t с $t_{\alpha;\,(n-2)}$. В этом случае, если t больше табличного, нулевая гипотеза отвергается.

При исследовании ранговой корреляции между двумя порядковыми переменными данные, необходимые для вычисления значения r_s, удобно представить в виде следующей таблицы (табл. 20).

Очень часто для ранжирования количественной переменной наблюдения располагают в виде вариационного ряда. При этом в качестве ранга берут номер наблюдения в данном ряду. В том случае, когда рядом оказываются равные значения переменной, то каждому из них присваивается среднее значение ранга. Например, имеются следующие значения переменной X: 11; 9; 15; 17; 9; 14; 18; 17. Располагаем их в вариационный ряд: 9; 9; 11; 14; 15; 17; 17; 18. Этой последовательности присваиваем следующие ранги: 1,5; 1,5; 3; 4; 5; 6,5; 6,5; 8. Для вычисления величины коэффициента ранговой корреляции Спирмэна используется программа П 29/34.

Таблица 20

Порядковый номер объекта		1	2	....	n
Ранги	Первая переменная	P_{11}	P_{12}	....	P_{1n}
	Вторая переменная	P_{21}	P_{22}	....	P_{2n}

Таблица 21

Студенты	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Вступит. экзамены	5	7	10	14	1	8	11	3	9	12	2	4	15	6	13
1-я сессия	2	5	8	6	4	7	9	10	11	12	1	3	13	15	14

Пример. При исследовании эффективности системы отбора абитуриентов в вузы города были обследованы несколько институтов. Для оценки эффективности производилось ранжирование подготовленности абитуриентов на вступительных экзаменах и по результатам первой экзаменационной сессии. Ниже приведены данные для одной из групп в количестве 15 человек (табл. 21).

Требуется определить величину коэффициента ранговой корреляции Спирмэна r_s.

Решение. Согласно инструкции к П 29/34 вводим программу и начинаем ввод данных таблицы: 5↑2 С/П (1) 7↑5 С/П (2) и т. д. 13↑14 С/П (15). По окончании ввода таблицы выполняем $\overrightarrow{\text{ШГ}}$ С/П ($2{,}5576112 = t$), далее ИП 9 ($5{,}785714 \cdot 10^{-1} = r_s$). По таблице находим значение t-критерия Стьюдента $t_{0{,}05;\ 13} = 1{,}77$. Поскольку $t > 1{,}77$, то можно говорить о наличии положительной корреляции.

К достоинствам коэффициента ранговой корреляции Спирмэна, кроме простоты его вычисления, относится и то, что это непараметрическая статистика. Однако следует помнить, что вследствие этого происходит и некоторая потеря информации. Значение r_s тем ближе к значению r_{xY}, чем сильнее линейность, корреляционной связи между изучаемыми признаками. Хотя метод ранговой корреляции и не требует обязательного наличия линейной корреляции между переменными, тем не менее необходимо, чтобы эта связь была монотонной.

Достоинства метода ранговой корреляции привели к тому, что этот метод широко используется в социологических, психологических, педагогических и ряде других исследований.

5.3. Таблицы сопряженности признаков и коэффициент ассоциации фи (φ)

В том случае, когда обе переменные измеряются в дихотомических шкалах наименований, в качестве меры связи используется коэффициент корреляции для дихотомических переменных — коэффициент φ (фи), называемый также коэффициентом ассоциации [23].

Рассмотрим следующий пример с использованием коэффициента φ. Для изучения свойств новой марки керамзитобетона были испытаны на сопротивление сжатию 45 проб бетона. При этом проверялось, выдержит ли пробный образец требуемое давление 4500 Н/см^2 или нет.

Пробы изготавливались по различной технологии: 25 образцов изготовили с применением «омагниченной» воды (прошедшей обработку в магнитном поле), а 20 образцов — по обычной технологии. Таким образом, переменная X принимала значения 1 (использование «омагниченной» воды) и 0 (обычная технология). Переменная Y принимала значения 1 (образец выдержал давление 450 кГ/см^2) и 0 (образец не выдержал давление 4500 Н/см^2). Значение коэффициента φ вычисляется по формуле

$$\varphi = \frac{P_{XY} - P_X \cdot P_Y}{\sqrt{P_X (1 - P_X) P_Y (1 - P_Y)}}, \qquad (5.3.1)$$

где P_X — доля «1» в массиве X, P_Y — доля «1» в массиве Y, P_{XY} — доля «1» одновременно в массиве X и Y. Согласно условиям примера $P_X = 0{,}5556$. Известно, что из 45 проб испытания выдержали 33 образца, т. е. $P_Y = 0{,}7333$, в том числе 21 образец с использованием «омагниченной» воды, т. е. $P_{XY} = 0{,}4667$. Для вычисления коэффициента корреляции Пирсона для дихотомических переменных используется программа П 30/34.

Согласно инструкции к программе П 30/34 вводим данные рассмотренной задачи и получаем значение $\varphi = 2{,}6975957 \cdot 10^{-1}$. В том случае, если оперируют не долями, а располагают абсолютные значения дихотомических данных в таблице сопряженности признаков (таблице, показывающей совместное появление пар значений по двум переменным в группе), то коэффициент Пирсона φ вычисляют по формуле

$$\varphi = \frac{b \cdot c - a \cdot d}{\sqrt{(a + c) \cdot (b + d) \cdot (a + d) \cdot (c + d)}}, \qquad (5.3.2)$$

где значения a, b, c и d приведены в следующей таблице сопряженности признаков 2×2 (табл. 22).

Выражение (5.3.2) было приведено Карлом Пирсоном в 1901 году в статье, опубликованной в «Философ-

Таблица 22

Признак Y	Признак X		Итого
	1	0	
1	a	в	a+в
0	c	d	c+d
Итого...	a+c	в+d	n

ских трудах Лондонского королевского общества» и посвященной корреляции между качественными признаками. Для рассмотренного выше примера $n=45$, $a=21$; $b=12$; $c=4$; $d=8$. Для проверки значимости полученного значения коэффициента используют выражение

$$\chi^2_{расч}=n\cdot\varphi^2, \qquad (5.3.3)$$

где n — сумма всех частот. Величина $\chi^2_{расч}$, полученная по (5.3.3), имеет χ^2-распределение с одной степенью свободы для четырехклеточной таблицы сопряженности признаков. В том случае, если $\chi^2_{расч}>\chi^2_{\alpha;1}$, то гипотеза $H_0:\varphi=0$ отвергается, т. е. имеет наличие достоверная связь между дихотомическими переменными X и Y.

Программу П 31/34 применяют для вычисления коэффициента корреляции φ и значения величины $\chi^2_{расч}$ при использовании таблицы сопряженности 2×2.

Согласно инструкции к П 31/34 вводим программу и исходные данные задачи: 21 П6 12 П7 4 П8 8 П9. Вычисляем значение φ : В/О С/П ($2{,}6967995\cdot10^{-1}$). Разница в четвертом знаке между этим значением φ и предыдущим, вычисленным по программе П 30/34, вызвана округлением значений P_{XY}, P_X и P_Y. Далее вычисляем значение $\chi^2_{расч}$: С/П ($3{,}2727274=\chi^2_{расч}$). Проверим нулевую гипотезу $H_0:\varphi=0$. По таблице [17, 18] находим критическое значение статистики χ^2 при $f=1$ и $\alpha=0{,}05$; $\chi^2_{1;\,0{,}05}=3{,}84$. Поскольку $\chi^2_{расч}$ меньше табличного значения, то у нас нет оснований отвергнуть нулевую гипотезу и мы ее принимаем. Это означает, что в используемой технологии производства керамзитобетона и «омаг-

ничивания» воды замена обычной воды на «омагниченную» не приводит к достоверному, статистически значимому увеличению прочности бетона.

5.4. Критерий случайной последовательности для дихотомической переменной

При работе с дихотомическими переменными нередко требуется произвести проверку последовательности значений этой переменной на случайное чередование этих значений. Например, при изготовлении образцов керамзитобетона предпочтительнее производить изготовление образцов с обычной и «омагниченной» водой в случайной последовательности, т. к. это повышает надежность результатов анализа и уменьшает влияние временнóго дрейфа технологии.

Для проверки массива, состоящего из последовательности дихотомических данных, например 0 и 1, на случайность чередования этих значений используется критерий серий Вальда-Вольфовитца, который вычисляется по следующей формуле [25]:

$$z = \frac{\left[P - \left(\frac{2n_1 \cdot n_0}{n_1 + n_0} + 1 \right) \right]}{\left[\frac{2n_1 \cdot n_0 (2n_1 \cdot n_0 - n_1 - n_0)}{(n_1 + n_0 - 1) \cdot (n_1 + n_0)^2} \right]^{1/2}}, \qquad (5.4.1)$$

где n_0 — количество «0» в последовательности; n_1 — количество «1» в последовательности; P — количество серий в последовательности. Рассмотрим последовательность **11 000 1 0 1 00 11 0 111 0 1.** В этой последовательности 11 серий (в каждой серии находятся либо только нули, либо только единицы), т. е. $p=11$, $n_0=8$ и $n_1=10$. Следует отметить, что выражение (5.4.1) используют при n_1 и n_0 более 20. Для проверки значимости полученного значения z используют следующие неравенства:

$$|z| < 1{,}96 \ (\alpha = 5\%) \quad \text{и} \quad |z| < 2{,}58 \ (\alpha = 1\%).$$

При выполнении этих неравенств порядок элементов в массиве случаен. Программа П 32/34 служит для вычисления значения z согласно (5.4.1).

5.5. Точечно-бисериальный коэффициент корреляции Пирсона

Довольно часто в практике исследований, особенно биологических и социально-экономических, возникает ситуация, когда одна переменная дихотомическая, а другая переменная измерена в шкале отношений или шкале интервалов. Например, переменная X обладает только альтернативной вариацией, представленной в форме «женский пол — мужской пол», «да — нет», «согласен— не согласен», «за — против» и т. п., а переменная Y измерена количественно в одной из следующих размерностей: см, кг, см/с, руб, с и т. п. Связь между переменными X и Y в таком случае называют двухстрочечной, или точечно-бисериальной корреляцией. Слово «точечно» подчеркивает, что признак X дискретен и имеет только 2 значения. Если же принимается, что признак X является непрерывным признаком, например, между ответами «согласен — не согласен» может существовать еще какая-то шкала, которая для простоты представлена в виде альтернативных ответов, то говорят о бисериальной корреляции. Поэтому необходимо отличать точечно-бисериальную корреляцию от просто бисериальной корреляции.

Вычисление точечно-бисериального коэффициента корреляции Пирсона производится с помощью программы П 33/34. Проверка значимости бисериальных коэффициентов корреляции производится таким же образом, что и проверка значимости парных линейных коэффициентов корреляции r_{xY}, т. е. с помощью t-статистики, либо сравнением вычисленного значения $r_{т.\,бис}$ с табличным критическим значением для объема наблюдений n и уровня значимости α.

ЗАДАЧИ К ГЛ. 5

1 [22, с. 213]. Ниже в таблице приведены ранги 15 студентов за ответы на два экзаменационных вопроса. Проверить с помощью коэффициентов ранговой корреляции, имеется ли корреляция между оценками на 1-й и 2-й вопросы. Значение $\alpha=5\%$ (табл. 23).

2 [46, с. 181]. В выборке из 100 кроликов исследовалась связь между окраской шерсти и цветом глаз. Результаты приведены в таблице 2×2 (табл. 24).

Вычислить коэффициент Пирсона φ и проверить его значимость при $\alpha=5\%$.

3 [22, с. 154]. В статье по генетике Батесон и Паннетт регистрировали цвет и форму 427 частиц пыльцы. Эти данные приведены в табл. 25. Позволяют ли они предположить, что зависимость между цветом и формой частиц существует? Уровень значимости $\alpha=5\%$.

4 [42, с. 361]. При измерении длины головы (X) и длины грудного плавника (Y) у 16 окуней получены следующие результаты (в мм).

Найти коэффициент ранговой корреляции и проверить его значимость при $\alpha=5\%$ (табл. 26).

5 [42, с. 361]. Связь между массой тела (X) и количеством гемоглобина в крови (Y) у павианов-гамадрилов характеризуется следующими данными (табл. 27).

Найти коэффициент ранговой корреляции.

6. При подбрасывании монеты 49 раз последовательность результатов имела следующий вид: Р ГГГ Р Г РРРР ГГ Р ГГГ РРРР Г Р Г РРР Г Р ГГГ РРР ГГГГ РРР Г Р ГРРР Г Р. Здесь Г — выпадение герба. Р — выпадение решетки. Является ли полученная последовательность случайной?

РЕШЕНИЯ

1. После набора программы П 29/34 начинаем ввод таблицы: 15↑15 С/П (1) и т. д. ... 3↑3 С/П (15). Далее $\vec{\text{ШГ}}$ С/П ($3{,}5697649=t$) ИП9 ($7{,}035714\cdot10^{-1}=r_s$). Поскольку $t>1{,}77=t_{0{,}05;\,13}$, то можно принять с $\alpha=5\%$ гипотезу о положительной корреляции между оценками за 1-й и 2-й вопросы на экзамене.

2. После ввода программы П 30/34 вводим исходные данные: 29 П6 11 П7 1 П8 59 П9 В/О С/П ($7{,}5724019\cdot10^{-1}=\varphi$). Далее С/П ($57{,}341271=\chi^2$). По таблице [17,

Таблица 23

Студент	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Вопрос 1, ранг	15	12,5	8	4	1	12,5	2	8	8	5	6	10,5	14	10,5	3
Вопрос 2, ранг	15	2	10	7,5	1	10	5,5	7,5	10	4	5,5	14	13	12	3

Таблица 24

Окраска шерсти	Окраска глаз		Σ
	красные	некрасные	
Белая	29	11	40
Окрашенная	1	59	60
Σ	30	70	100

Таблица 25

Форма	Цвет		Всего
	пурпурный	красный	
Вытянутые	296	27	323
Круглые	19	85	104
Всего...	315	112	427

Таблица 26

X	66	61	67	73	51	59	48	47	58	44	41	54	52	47	51	45
Y	38	31	36	43	29	33	28	25	36	26	21	30	20	27	28	26

Таблица 27

Масса тела, кг	18	17,7	19	18	19	22	21	21	20	30
Гемоглобин (по Сали)	70	74	72	80	77	80	80	89	76	89

Таблица 28

41	44	45	47	47	48	51	51	52	54	58	59	61	66	67	73
21	26	26	27	25	28	29	28	20	30	36	33	31	38	36	43

Таблица 29

1	2	3	4,5	4,5	6	7,5	7,5	9	10	11	12	13	14	15	16
2	4,5	4,5	6	3	7,5	9	7,5	1	10	13,5	12	11	15	13,5	16

18] находим $\chi^2_{0,05;\,1}=3,84$. Поскольку $\chi^2>3,84$, то принимается гипотеза о связи между окраской шерсти у кроликов и цветом глаз.

3. Используя программу П 31/34, получаем следующее решение: $\varphi=7,1594859\cdot10^{-1}$; $\chi^2=218,87268$. По таблице [17, 18] находим $\chi^2_{0,05;\,1}=3,84<\chi^2=218,87268$, поэтому отвергаем нулевую гипотезу и принимаем альтернативную: есть свидетельство в пользу того, что зависимость между цветом и формой частиц пыльцы существует.

4. Расположив в порядке возрастания длины головы (X), получим следующую таблицу (табл. 28).

Далее, проранжировав X и Y, приходим к следующим результатам (табл. 29).

Обработав эти результаты с помощью программы П 29/34, получим $r_s=8,588235\cdot10^{-1}$, $t=6,2728895$. По таблице [17, 18] находим $t_{0,05;\,14}<6,27$, т. е. корреляция значима.

5. Используя ранжирование X и Y, с помощью программы П 29/34 находим $r_s=7,060606\cdot10^{-1}$, $t=2,8200763$.

6. Определяем характеристики последовательности, число серий $P=25$; число гербов $n_2=22$, число решеток $n_p=27$. После набора программы П 32/34 вводим эти числа в ПМК 34 и производим вычисление. В результате получим $z=-7,1471038\cdot10^{-2}$, поскольку $|z|<1,96$, то принимаем нулевую гипотезу о случайном характере последовательности.

ГЛАВА 6

ОСНОВЫ ДИСПЕРСИОННОГО АНАЛИЗА НАБЛЮДЕНИЙ

Различие взглядов зависит еще от способа определения влияния известных нам данных.

П. С. Лаплас

6.1. Основные идеи дисперсионного анализа

Целью дисперсионного анализа является расщепление суммарной дисперсии изучаемой величины на отдельные составляющие: дисперсию, обусловленную действием одного или нескольких факторов, изменяющих свое значение от эксперимента к эксперименту, и дисперсию, вызванную совместным действием всех остальных неучитываемых факторов. Из этого определения следует, что дисперсионный анализ можно применять для изучения явлений, процессов, которые описываются или характеризуются минимум двумя характеристиками. Одна из этих характеристик является составляющей вектора выходных параметров процесса Y_i, а другая— составляющей вектора входных параметров X_j. При этом характеристика Y_i обязательно должна быть количественной: мощность, концентрация вещества, скорость, активность, плотность, урожайность, привес животных, производительность за смену и т. д. Другая же характеристика X_j, которую можно целенаправленно изменять, может быть как количественной, так и качественной. Примеры количественной характеристики X_j: температура реакций, расход технологического газа, давление, содержание тех или иных компонент, время закалки или отжига и т. д. Примеры качественных характеристик переменной X_j: номер смены на производстве, номер установки или прибора, квалификация рабочего, номер диеты или рациона кормления, наименование органов в теле человека или животного, пол, образование и т. д.

Основные схемы использования дисперсионного анализа и его идеи проиллюстрируем следующим примером. Катализатор для химической реакции получали в семи разных режимах, отличавшихся температурой и

временем спекания. Всего было проведено в общей сложности 35 процессов получения катализатора. После получения катализатора измерялась активность катализатора у всех 35 образцов. Необходимо было установить, влияет ли режим получения катализатора на его активность, которая для 35 образцов изменялась от минимальной в 20 единиц до максимальной в 60 единиц. После обработки результатов измерения активности было установлено, что средняя активность в выборке из 35 образцов равна 35 единицам, а стандартное отклонение равно 7 единицам, т. е. $s^2=49$. Принимая значение $s^2=49$ за 100%, можно с помощью дисперсионного анализа установить, что, например, 90% этой суммарной дисперсии активности обусловлены, связаны с изменением режима получения катализатора (7 режимов) и только 10% — всеми прочими колебаниями других технологических параметров. При таком соотношении 90 : 10 можно говорить о доминирующем влиянии фактора — режима получения катализатора, на его активность по сравнению с колебаниями всех остальных параметров. При обратном же соотношении 10 : 90 этого утверждать уже нельзя. В этом случае для управления активностью получаемого катализатора необходимо будет выявить те основные факторы (один или несколько), которые связаны с этими необъясненными 90%.

Таким образом, первая схема — расщепление суммарной дисперсии количественного признака Y_i на составляющие части, которые связаны с контролируемым и изменяемым фактором X_j (или несколькими такими факторами) и остальными неконтролируемыми и неуправляемыми факторами. Анализ расщепленной на составные части дисперсии позволяет ответить на вопрос о значимости влияния фактора X_j на величину признака Y_i.

Разработка методов дисперсионного анализа связана с именем известного английского статистика Р. А. Фишера. Эти работы были начаты им на Ротэмстедской агробиологической станции, основанной в 1920 г. Целью этих работ было изучение вклада агробиологических факторов, таких как сорт семян, удобрения, условия роста растений и т. д., на урожайность и другие характеристики растений.

Рассмотренную выше задачу о 7 режимах получения катализаторов можно представить несколько иначе. В каждом из 7 режимов было получено по 5 образцов катализатора, при этом после первичной обработки результатов измерения активности было получено 7 значений средней активности катализатора в каждой группе: $\bar{Y}_1$; $\bar{Y}_2$; $\bar{Y}_3$; $\bar{Y}_4$; $\bar{Y}_5$; $\bar{Y}_6$; $\bar{Y}_7$. В главе 4 мы рассмотрели методы сравнения двух выборочных средних. Если бы режимов получения катализатора было не 7, а 2, тогда, сравнив $\bar{Y}_1$ и $\bar{Y}_2$ между собой, можно было бы ответить на вопрос: влияет ли изменение режима получения катализатора на среднюю активность или не влияет? Если бы нулевая гипотеза $H_0: \mu_1 = \mu_2$ была принята, то это означало бы, что имеющиеся различия между первым и вторым режимом не привели к значимому изменению средней активности катализатора, и наоборот. Но в данном случае мы имеем не 2, а 7 режимов. Как быть? Сравнивать их попарно: $\bar{Y}_1$ и $\bar{Y}_2$; $\bar{Y}_2$ и $\bar{Y}_3$ и т. д.? Но такая проверка не гарантирует надежного заключения. Решение этой задачи сравнения нескольких средних также обеспечивает дисперсионный анализ. Нетрудно заметить, что эта вторая схема дисперсионного анализа фактически идентична первой схеме.

Мы рассмотрим основы только однофакторного дисперсионного анализа. Для более полного изучения теории дисперсионного анализа рекомендуем читателю специальную литературу [2, 3, 19]. После этого будут даны программы для однофакторного и двухфакторного дисперсионного анализа с учетом эффекта взаимодействия. На примерах этих программ мы продемонстрируем взаимосвязь первой и второй постановки исходной задачи, т. е. расщепления дисперсии на составляющие части и сравнения нескольких средних.

6.2. Однофакторный дисперсионный анализ

6.2.1. Анализ сгруппированных данных

Наиболее удобно представлять данные для однофакторного дисперсионного анализа в виде таблицы (табл. 30). Каждый столбец такой таблицы содержит

одну выборку наблюдений, полученных при постоянном значении фактора X. Каждое наблюдение обозначается Y_{ij}, где индекс i указывает номер этого наблюдения (или строки) в столбце с номером j.

Таблица 30

Номер наблюдений (строки) i	Номер режима (столбца) j				
	1	2	3	j	K
1	Y_{11}	Y_{12}	Y_{13}	...	$Y_{1к}$
2	Y_{21}	Y_{22}	Y_{23}	...	$Y_{2к}$
3	Y_{31}	Y_{32}	Y_{33}	...	$Y_{3к}$
	...	...	...	...	...
i	Y_{i1}	Y_{i2}	Y_{i3}	Y_{ij}	$Y_{iк}$
⋮	...	...	...	...	...
n	Y_{n1}	Y_{n2}	Y_{n3}	Y_{nj}	$Y_{nк}$
Среднее по режиму (столбцу)	$\overline{Y}_{\cdot 1}$	$\overline{Y}_{\cdot 2}$	$\overline{Y}_{\cdot 3}$	$\overline{Y}_{\cdot j}$	$\overline{Y}_{\cdot к}$
Общее среднее по всей серии наблюдений					$\overline{Y}_{\cdot\cdot}$

В рассмотренной выше задаче индекс j принимал значения от 1 до 7, а индекс i от 1 до 5. В задачах, решаемых методом дисперсионного анализа, равное число наблюдений встречается не всегда, поэтому рассмотрим более общий случай, допуская неравное число наблюдений в отдельных столбцах.

Для упрощения вычислений, которые потребуется далее выполнять, возьмем таблицу с тремя разными режимами спекания катализатора и неравным числом экспериментов для каждого режима (табл. 31).

Для большего удобства обозначений примем

$$Y_{\cdot j} = \sum_{i=1}^{n_j} Y_{ij}, \quad j = 1, 2, \ldots, k, \qquad (6.2.1)$$

т. е. для любого столбца (режима) сумма значений равна $Y_{\cdot j}$, число наблюдений в столбце, т. е. число экспериментов при значении входной переменной X_j, обозначено как n_j. Двойной индекс $\cdot j$ обозначает, что суммирование ведется по всем наблюдениям j-того режима. Сум-

Таблица 31

Номер эксперимента в режиме	Номер режима		
	1	2	3
1	10	12	15
2	12	25	18
3	18	13	45
4	8	20	25
5	16	30	12
6	30		18
7	11		35
Средняя активность катализатора в режиме	15	20	24
Общая средняя активность катализатора во всей серии экспериментов	19,631 57		

му всех наблюдений серии из j режимов обозначим через

$$Y_{..} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij}. \qquad (6.2.2)$$

В выражении (6.2.2) индекс $\cdot\cdot$ указывает, что суммирование ведется как по всем строкам, так и по всем столбцам (режимам).

Средние значения величины Y по j-тому режиму (столбцу) обозначим через

$$\bar{Y}_{\cdot j} = \frac{Y_{\cdot j}}{n_j}. \qquad (6.2.3)$$

Общее среднее по всей серии экспериментов обозначим через

$$\bar{Y}_{..} = \frac{Y_{..}}{N}, \qquad (6.2.4)$$

где общее число экспериментов $N = n_1 + n_2 + ... + n_k$. В том случае, когда в каждом режиме, т. е. при $X_j = const$, проводили одинаковое число экспериментов n, тогда $N = n \cdot k$.

Для каждого значения Y_{ij} имеет место следующее равенство:

$$(Y_{ij} - \bar{Y}_{..}) = (\bar{Y}_{\cdot j} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{\cdot j}) . \quad (6.2.5)$$

Выражение (6.2.5) расщепляет отклонение любого Y_{ij} от общего среднего $\bar{Y}_{..}$ на 2 слагаемых. Первое слагаемое представляет собой отклонение среднего по столбцу от общего среднего — отклонение столбца. Второе слагаемое представляет собой отклонение Y_{ij} от среднего по своему столбцу — остаточное отклонение. В словесной формулировке выражение (6.2.5) можно записать так: общее отклонение=отклонение столбца + остаточное отклонение.

Левую и правую части выражения (6.2.5) возведем в квадраты и просуммируем по всем N наблюдениям. При этом отклонения преобразуются в меры вариации

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{j=1}^{k} n_j (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2 +$$

$$+ \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\cdot j})^2 . \quad (6.2.6)$$

В словесной формулировке выражение (6.2.6) имеет вид общая вариация=вариация столбцов + остаточная вариация. Это соотношение является основным для понимания дальнейших выкладок.

Предположим далее, что для каждого режима (столбца) значение можно представить в виде следующей суммы:

$$Y_{ij} = \mu_{\cdot j} + \varepsilon_{ij}, \quad (6.2.7)$$

где $\mu_{\cdot j}$ — генеральное среднее значение величины Y для режима j, т. е. при $X = X_j$, а ε_{ij} — ошибка с нулевым генеральным средним и дисперсией σ^2. Предполагается, что ошибки ε_{ij} имеют нормальное распределение $N_1(0; \sigma^2)$ и не взаимосвязаны друг с другом. Аналогичным образом среднее по каждому столбцу (режиму) представим в виде

$$\mu_{\cdot j} = \mu_{..} + \tau_j, \quad j = 1, 2, ..., k, \quad (6.2.8)$$

где $\mu_{..}$ — генеральное среднее всей совокупности экспериментов с разными значениями X_j, а величина τ_j называется эффектом столбца. По выражению (6.2.8) $\tau_j = \mu_{\cdot j} - \mu_{..}$, т. е. эффект столбца равен разности между

средним по режиму и общим средним. Сумма же подобных эффектов столбцов, т. е. $\sum_{j=1}^{k} \tau_j$, в генеральной совокупности предполагается равной нулю. Это означает, что общее среднее есть среднее арифметическое всех средних по отдельным столбцам. Подставляя выражение (6.2.8) в (6.2.7), получим

$$Y_{ij} = \mu_{..} + \tau_j + \varepsilon_{ij}. \qquad (6.2.9)$$

В рамках принятой модели оценки для генеральных средних по столбцам $\mu_{.j}$, вычисляемые по методу наименьших квадратов, равны

$$\hat{\mu}_{\cdot j} = \bar{Y}_{\cdot j}, \qquad (6.2.10)$$

где знак «^» заменяет слово «оценка».

Тогда, подставляя (6.2.10) в (6.2.7), имеем

$$Y_{ij} = \bar{Y}_{\cdot j} + e_{ij}, \; j = 1,2,\ldots, k. \qquad (6.2.11)$$

где e_{ij} — остаток наименьших квадратов. В этом слячае остаточная сумма квадратов для j-того столбца имеет следующий вид:

$$\sum_{i=1}^{n_j} e^2_{ij} = \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\cdot j})^2 \qquad (6.2.12)$$

с числом степеней свободы $f = n_j - 1$, т. к. среднее $\bar{Y}_{.j}$ оценивается по n_j наблюдениям. Для всех столбцов и строк общую остаточную сумму запишем так:

$$S_1 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} e^2_{ij} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\cdot j})^2 \qquad (6.2.13)$$

с числом степеней свободы $N - k = \sum_{j=1}^{k} (n_j - 1)$. Сравнивая (6.2.13) и (6.2.6), нетрудно заметить, что S_1 — это остаточная вариация.

Рассмотрим далее, каким же образом все эти суммы и суммы квадратов могут помочь нам в решении задачи, например, в проверке нулевой гипотезы о равенстве средних по столбцам — режимам

$$H_0: \mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot k} = \mu_{..}$$

против альтернативной гипотезы H_1: не все средние по столбцам равны.

Эта гипотеза эквивалентна гипотезе о том, что эффекты от изменения значения входной переменной X, т. е. эффекты столбцов τ_j, равны нулю против альтернативной гипотезы о том, что не все τ_j равны нулю. Если нулевая гипотеза верна, тогда в выражении (6.2.7) можно каждое $\mu_{\cdot j}$ заменить на $\mu_{\cdot\cdot}$, т. е.

$$Y_{ij} = \mu_{\cdot\cdot} + e_{ij}, \qquad (6.2.14)$$

а также заменить $\hat{\mu}_{\cdot\cdot}$ на $\bar{Y}_{\cdot\cdot}$, и тогда

$$Y_{ij} = \bar{Y}_{\cdot\cdot} + e_{ij}. \qquad (6.2.15)$$

Итак, если нулевая гипотеза действительно верна, то тогда общая остаточная сумма квадратов в (6.2.15) запишется так:

$$S_o = \sum_{j=1}^{k} \sum_{i=1}^{n_j} e^2_{ij} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 \qquad (6.2.16)$$

с N—1 степенями свободы. Согласно (6.2.6) S_o назовем остаточной вариацией.

Используя данные табл. 14, вычислим необходимые нам далее значения отдельных сумм.

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij}^2 = 10^2 + 12^2 + 18^2 + \ldots + 12^2 + 25^2 +$$
$$+ 13^2 + \ldots + 15^2 + 18^2 + \ldots + 35^2 =$$
$$= 1909 + 2238 + 4892 = 9039.$$

Общую вариацию легче найти, преобразовав выражение (6.2.6) к следующему виду:

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\ldots})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} Y^2_{ij} - \frac{\left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij}\right)^2}{N}. \qquad (6.2.17)$$

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij} = 10 + 12 + 18 + 8 + \ldots + 12 + 25 + \ldots +$$
$$+ 15 + 18 + \ldots + 35 = 373.$$

Тогда имеем

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = 9039 - \frac{373^2}{19} = 1716{,}421 = S_o.$$

Вычислим вариацию столбцов

$$\sum_{j=1}^{k} n_j(\bar{Y}_{.j}-\bar{Y}_{..})^2=7\cdot(15-19{,}63157)^2+5\cdot(20-19{,}63157)^2+$$
$$+7\cdot(24-19{,}63157)^2=284{,}421=S_c.$$

Остаточная вариация равна разности общей вариации и вариации столбцов

$$S_1=1716:421-284{,}421=1432=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j})^2.$$

В том случае, если нулевая гипотеза верна, то общая вариация S_0 и остаточная вариация S_1 будут близки по величине, а их разность — вариация столбцов S_c, будет незначима и достаточно мала по сравнению с вариацией S_1. Исходя из этих соображений и строится критерий проверки нулевой гипотезы, так называемое F-отношение:

$$F_{расч}=\frac{S_c/(K-1)}{S_1/(N-K)}=\frac{(\text{вариация режимов})/(K-1)}{(\text{остаточная вариация})/(N-K)}. \qquad (6.2.18)$$

Данный критерий для проверки нулевой гипотезы подчиняется F-распределению Фишера-Снедекора с (K—1) и (N—K) степенями свободы. Если величина F-отношения, вычисленная согласно (6.2.18), больше критического значения F из таблицы F-распределения $F_{\alpha;\, f_1;\, f_2}$, то нулевая гипотеза отвергается.

Результаты вычислений при дисперсионном анализе удобно оформлять в виде таблицы (табл. 32). Средний квадрат остаточной вариации является оценкой дисперсии ошибки ε_{ij}, т. е. $\hat{\sigma}^2$. Если нулевая гипотеза H_0 принята, то и средний квадрат вариации столбцов также является оценкой дисперсии $\hat{\sigma}^2$.

Определим величину F-отношения для нашего примера

$$F=\frac{284{,}421/2}{1432/(19-3)}=1{,}59,$$

Принимая $\alpha=5\%$ по таблице F-распределения с числом степеней свободы $f_1=2$ и $f_2=16$, находим критическое

значение критерия, равное 3,63. Поскольку вычисленное значение F-отношения меньше критического, то при заданном α нулевая гипотеза принимается. Итак, несмотря на то, что, на первый взгляд, мы имеем разные средние активности катализатора для трех разных режимов,

Таблица 32

Источник вариации	Величина вариации	Число степеней свободы	Средний квадрат	F-отношение
Общая	$s_0=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{..})^2$	N—1		$F=\frac{s_c/(k-1)}{s_1/(N-k)}$
Столбцы (режимы)	$s_c=\sum_{i=1}^{k}(\bar{Y}_{.j}-\bar{Y}_{..})^2$	k—1	$\frac{s_c}{k-1}$	
Ошибка (влияние неучтенных факторов)	$s_1=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j})^2$	N—k	$\frac{s_1}{N-k}$	

нельзя считать эти различия средних не случайными, значимыми, достоверными. Факт принятия нулевой гипотезы о равенстве трех генеральных средних активностей необходимо понимать следующим образом: изменение температуры и времени спекания в имеющемся диапазоне значений не оказывает значимого изменения на генеральную среднюю активность катализатора в отдельном режиме. Из полученных результатов имеем оценку $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = s^2 = \frac{S_1}{N-k} = \frac{1432}{16} = 89{,}5\,,$$

т. е. $s=9{,}46$.

Не следует после принятия нулевой гипотезы делать категорический вывод о том, что варьируемый от режима к режиму входной параметр процесса X вообще не может оказывать воздействия на выходной пара-

метр Y. Вполне возможно, что эти два параметра могут быть связаны между собой определенной зависимостью. Однако по двум возможным причинам эта связь не выявлена. 1. Диапазон изменения параметра X очень мал и находится в пределах точности поддержания этого параметра, т. е. мы предполагаем, что изменили значение X, тогда как фактически оно осталось прежним. 2. Точность поддержания всех остальных входных параметров недостаточна, вследствие этого высокое значение «шума» ($s^2=89,5$) «забивает» влияние варьирования X.

Для проверки первой причины необходимо расширить диапазон изменения параметра X, естественно, в разумных, рабочих пределах.

Для проверки второй причины, если имеются серьезные основания (и необходимость) для опровержения нулевой гипотезы, необходимо попытаться повысить стабилизацию остальных входных параметров либо внимательно изучить другие возможные причины невоспроизводимости при постоянном значении X.

6.2.2. Применение линейных контрастов

Кроме Р. Фишера методу дисперсионного анализа много внимания уделил другой известный статистик Г. Шеффе. Именно он предложил оригинальный метод линейных контрастов, называемый в литературе S-методом (от Scheffe H.). Рассмотрим идею этого метода и его реализацию в том случае, когда нулевая гипотеза отвергнута.

Напомним, что нулевая гипотеза формулируется как предположение о равенстве генеральных средних каждого режима между собой и общему генеральному среднему. Однако рассмотренная выше процедура дисперсионного анализа не дает информации о том, *какие именно средние* привели к отбрасыванию нулевой гипотезы, если окажется, что вычисленное F-отношение больше критического. Вполне возможно, что *некоторые средние равны* между собой, а некоторые — *нет*.

Линейный контраст $\hat{L}$ представляет собой обычную линейную комбинацию средних по столбцам

$$\hat{L} = \sum_{j=1}^{k} C_j \bar{Y}_{\cdot j}, \qquad (6.2.19)$$

причем обязательно должно выполняться условие

$$\sum_{j=1}^{k} C_j = 0. \qquad (6.2.20)$$

Оценка дисперсии линейного контраста $s^2_{\hat{L}}$ равна

$$s^2_{\hat{L}} + \hat{\sigma}^2 \cdot \sum_{j=1}^{k} \frac{C_j^2}{n_j}. \qquad (6.2.21)$$

Для линейного контраста $\hat{L}$ можно определить $100 \cdot (1-\alpha)\%$-ный доверительный интервал по формуле

$$\hat{L} \pm s_{\hat{L}} \cdot \sqrt{(k-1) \cdot F_{\alpha;\, k-1;\, N-k}}, \qquad (6.2.22)$$

где $\hat{L}$ находится согласно (6.2.19).

Предположим, что мы провели серию экспериментов по получению катализатора не с тремя разными режимами, а с четырьмя, расширив при этом диапазон изменения температуры и времени спекания. При этом средние по каждому отдельному режиму оказались следующие: $\bar{Y}_{\cdot 1}=13$; $\bar{Y}_{\cdot 2}=21$; $\bar{Y}_{\cdot 3}=25$; $\bar{Y}_{\cdot 4}=30$, а объемы выборок равны $n_1=5$; $n_2=7$; $n_3=6$; $n_4=4$. Новая оценка дисперсии $s^2=25{,}4$.

При проведении дисперсионного анализа нулевая гипотеза о равенстве всех 4 средних по режимам была отвергнута при $\alpha=5\%$. В данной задаче возможны шесть линейных контрастов, т. е. можно проверить предположения о следующих 6 возможных причинах отклонения нулевой гипотезы: $\mu_{\cdot 1} \neq \mu_{\cdot 2}$, $\mu_{\cdot 1} \neq \mu_{\cdot 3}$, $\mu_{\cdot 1} \neq \mu_{\cdot 4}$, $\mu_{\cdot 2} \neq \mu_{\cdot 3}$, $\mu_{\cdot 2} \neq \mu_{\cdot 4}$, $\mu_{\cdot 3} \neq \mu_{\cdot 4}$. Для проверки нулевой гипотезы $H_0: \mu_{\cdot 1} - \mu_{\cdot 2} = 0$ против двусторонней альтернативной гипотезы $H_1: \mu_{\cdot 1} \neq \mu_{\cdot 2}$ значения C_j в выражениях (6.2.19), (6.2.20) и (6.2.21) равны $C_1=1$; $C_2=-1$; $C_3=C_4=0$.

Вначале по (6.2.19) находим первый линейный контраст

$$\hat{L}_{12} = 1 \cdot 13 - 1 \cdot 21 = -8.$$

Далее по (6.2.21) определим среднеквадратическое отклонение $s_{\hat{L}}$ для этого контраста

$$s^2_{\hat{L}} = 25,4\left(\frac{1}{5} + \frac{1}{7}\right) = 8,708571, \quad s_{\hat{L}} = 2,95.$$

Наконец, согласно (6.2.22) находим доверительный интервал для $\hat{L}_{12}$, задавшись $\alpha = 5\%$ ($F_{0,05;\,3;\,18} = 3,16$): $-8 \pm 2,95 \cdot \sqrt{(14-1) \cdot 3,16} = -8 \pm 9,08$. Итак, согласно (6.2.22) имеем доверительный интервал

$$-8-9,08 \leqslant \mu_{\cdot 1} - \mu_{\cdot 2} \leqslant -8+9,08,$$

$$-17,08 \leqslant \mu_{\cdot 1} - \mu_{\cdot 2} \leqslant 1,08.$$

Так как этот доверительный интервал для разности генеральных средних двух режимов покрывает нуль, то можно сделать вывод о том, что $\mu_{\cdot 1}$ и $\mu_{\cdot 2}$ являются незначимо, несущественно различающимися. Проведя аналогичные вычисления для остальных пяти пар, получим следующие результаты:

$$\begin{aligned} &-21,4 \leqslant \mu_{\cdot 1} - \mu_{\cdot 3} \leqslant -2,6, \\ &-27,41 \leqslant \mu_{\cdot 1} - \mu_{\cdot 4} \leqslant -6,59, \\ &-12,63 \leqslant \mu_{\cdot 2} - \mu_{\cdot 3} \leqslant 4,63, \\ &-18,73 \leqslant \mu_{\cdot 2} - \mu_{\cdot 4} \leqslant 0,73, \\ &-15,22 \leqslant \mu_{\cdot 3} - \mu_{\cdot 4} \leqslant 5,02. \end{aligned} \tag{6.2.23}$$

Из этих неравенств видно, что для первого и третьего, первого и четвертого режимов доверительные интервалы для разности генеральных средних не покрывают нуль, следовательно, именно эти различия и привели к отклонению нулевой гипотезы. Итак, линейные, или, как их еще называют, ортогональные, контрасты Шеффе позволяют нам как бы «заглянуть за кулисы» принятого решения, выявить его более тонкие детали и аргументы. Сравнивая между собой неравенства (6.2.23), мы более наглядно можем представить себе тот порог чувствительности процесса к изменениям входного параметра X, превысив который мы значимо, достоверно изменяем ответную реакцию процесса на это возмущение.

Кроме S-метода множественных сравнений в практике широко используются и другие методы сравнений, например T-метод Тьюки. Для начинающих знакомство

с этими методами впервые можно порекомендовать доступное их изложение в [19].

Прежде чем перейти непосредственно к программам дисперсионного анализа, рассмотрим два новых понятия. Обратимся вновь к выражению (6.2.6). Очевидно, что при постоянной общей вариации S_0 увеличение вариации столбцов S_c приводит к уменьшению остаточной вариации S_1, т. к. $S_0=S_c+S_1$. Увеличение S_c означает усиление связи выходной переменной Y с входной переменной X. Для количественной меры этой силы связи независимо от ее вида (линейная или нелинейная) используют коэффициент детерминации

$$\hat{\eta}^2 = \frac{S_c}{S_0}. \qquad (6.2.24)$$

Коэффициент детерминации всегда положителен и не больше 1. При $\eta^2=1$ имеется точная функциональная зависимость $Y=f(X)$, а при $0<\eta^2<1$ наблюдается стохастическая связь, которая проявляется в виде определенной закономерности только в совокупности наблюдений. Используя $\hat{\eta}^2$, выражение (6.2.18) можно записать в виде

$$F_{расч} = \frac{\hat{\eta}^2/(k-1)}{(1-\hat{\eta}^2)/(N-k)}. \qquad (6.2.25)$$

Положительное значение корня квадратного из коэффициента детерминации

$$\hat{\eta} = \sqrt{\frac{S_c}{S_0}} \qquad (6.2.26)$$

называется корреляционным отношением (не путать с коэффициентом корреляции, о котором речь пойдет ниже). Очевидно, что для величины $\hat{\eta}^2$, которая является точечной оценкой генерального параметра η^2, можно построить, как и для других параметров, доверительный интервал. Чем у́же будет доверительный интервал, тем точнее найденная точечная оценка $\hat{\eta}^2$.

Программа П 34/34 позволяет выполнять полный однофакторный дисперсионный анализ сгруппированных данных.

В программе П 35/34 реализован метод линейных контрастов. Набор этой программы необходимо производить по окончании работы программы П 34/34, т. к. необходимые для линейных контрастов данные вычисляются программой П 34/34. Набор программы П 35/34 производить, не выключая питания ПМК 34 после окончания работы П 34/34, т. к. это приведет к стиранию необходимых данных в ПМК 34.

6.2.3. Анализ усредненных данных

Часто в распоряжении исследователя имеется набор результатов первичной обработки экспериментов по каждому режиму в отдельности. В этом случае процедура выполнения дисперсионного анализа ускоряется за счет уменьшения количества вводимых чисел. При использовании усредненных данных результаты экспериментов удобно оформить следующей таблицей (табл. 33).

Таблица 33

Параметры	Значения фактора X			
	X_1	X_2	...	X_j
$\overline{Y}_{\cdot j}$; s_j	$\overline{Y}_{\cdot 1}$; s_1	$\overline{Y}_{\cdot 2}$; s_1	...	$\overline{Y}_{\cdot j}$; s_j
Количество процессов	n_1	n_2	...	n_j

С помощью следующей программы П 36/34 также проводится однофакторный дисперсионный анализ только с использованием усредненных данных, т. е. когда для каждого режима известны $\overline{Y}_{\cdot j}$; s_j; n_j. Линейные контрасты в этом случае реализуются с помощью программы П 37/34.

6.3. Двухфакторный дисперсионный анализ

В том случае, когда целенаправленно изменяются два входных фактора, например фактор А и фактор В, и исследуется влияние этих изменений на выходной параметр Y, используется двухфакторный дисперсионный

анализ. В общем случае при многих факторах говорят о многофакторном дисперсионном анализе. Критерии проверки нулевой гипотезы строятся подобно тому, как это делалось в случае однофакторного анализа. Однако есть и некоторые отличия. В частности, при двухфакторном дисперсионном анализе общую дисперсию параметра Y можно расщепить двумя способами. В первом случае можно записать следующее равенство:

$$\eta^2_\Sigma = \eta^2_X + \eta^2_Z = \eta^2_A + \eta^2_B + \eta^2_Z = 100\%, \quad (6.3.1)$$

где η^2_X — квадрат корреляционного отношения, обусловленный изменением составляющих вектора X факторами A и B; η^2_A и η^2_B — аналогично предыдущему, но отдельно для каждого фактора; η^2_Z — квадрат корреляционного отношения, обусловленный всеми прочими неучитываемыми факторами.

Во втором же случае расщепление дисперсии и η^2_Σ производится иначе

$$\eta^2_\Sigma = \eta^2_X + \eta^2_Z = \eta^2_A + \eta^2_B + \eta^2_{AB} + \eta^2_Z = 100\%. \quad (6.3.2)$$

В этом равенстве появилось новое слагаемое η^2_{AB} — квадрат корреляционного отношения, обусловленный эффектом взаимодействия двух факторов A и B. В многофакторном дисперсионном анализе такие эффекты взаимодействия могут быть тройные, четвертные и т. д. Однако, как правило, большинство эффектов выше тройных малозначимы.

6.3.1. Анализ сгруппированных данных

Для удобства ввода данных сгруппируем результаты наблюдения за исследуемой величиной в табл. 34.

Таблица 34

Значение фактора A	Значение фактора B		
	B_1	B_2	B_3
A_1	Y_{111}; Y_{112}; ...; $Y_{11n_{11}}$	Y_{121}; Y_{122}; ...; $Y_{12n_{12}}$	Y_{131}; Y_{132}; ...; $Y_{13n_{13}}$
A_2	Y_{211}; Y_{212}; ...; $Y_{21n_{21}}$	Y_{221}; Y_{222}; ...; $Y_{22n_{22}}$	Y_{231}; Y_{232}; ...; Y_{23n_2}

Изменения факторов А и В не обязательно при этом должны быть количественными изменениями. Это могут быть разные смены, разные бригады, разные единицы однотипного оборудования, т. е. в качестве изменяющихся факторов могут выступать и качественные признаки, характеризующие наблюдаемый процесс. При этом количества наблюдений при различных комбинациях А и В не обязательно должны быть равны между собой. Число уровней, на которых могут изменяться факторы А и В, в общем случае выбирается самим исследователем. Однако в связи с ограничением по числу регистров памяти в ПМК 34 ниже приведены программы двухфакторного дисперсионного анализа, в которых число уровней фактора А ограничено границами $2 \div 9$, а число уровней фактора В—$2 \div 3$. В максимальных значениях это приводит к 27 отличающимся комбинациям (клеткам), что вполне достаточно для решения задачи выделения доминирующих факторов.

С помощью микрокалькуляторов типа ПМК 34 двухфакторный дисперсионный анализ производится с использованием одной из программ П 38/34 для сгруппированных данных и П 39/34 для усредненных данных. При учете эффекта взаимодействия факторов А и В линейные контрасты образовать трудно вследствие усложняющего влияния этого взаимодействия.

Итак, дисперсионный анализ позволяет нам проверить очень важные предположения относительно наличия (отсутствия) влияния тех или иных входных факторов на величину выходного параметра. Важной особенностью этого метода является его универсальность: его можно применять как при дискретных переменных X_i (номер смены, номер конструкции и т. п.), так и при непрерывных переменных X_i. В результате дисперсионного анализа расщепление общей дисперсии позволяет проранжировать по вкладу в эту дисперсию входные факторы X_i и произвести проверку значимости величин $\hat{\eta}^2_i$. При этом не ставится вопрос о том, как зависит параметр Y_j от параметра X_i, какого вида имеется связь: линейная или нелинейная и т. п. Однако для целей прогнозирования результатов и управления исследуемым процессом недостаточно только знания того, что эта зависимость имеется. Необходимо оценить и ее вид

(линейная или нелинейная), а также получить оценку для выражения, описывающего эту зависимость в виде уравнения с определенными числовыми коэффициентами, связывающего значения Y_j с X_i.

Именно эта задача и рассматривается далее в главе 7, где помимо кратких основ методов даны и реализующие их программы с примерами.

ЗАДАЧИ К ГЛ. 6

1 [40, с. 210]. Для проверки влияния методики обучения производственным навыкам на качество подготовки отбираются случайным образом из выпускников ПТУ четыре группы учеников, которые после обучения (по разным методикам) показали производственные показатели, приведенные в табл. 35.

Таблица 35

Номер ученика в группе	Выработка по группам учеников, детали/день			
	1-я группа	2-я группа	3-я группа	4-я группа
1	60	75	60	95
2	80	66	80	85
3	75	85	65	100
4	80	80	60	80
5	85	70	86	
6	70	80	75	
7		90		

Представьте теперь, что вы — директор ПТУ, и вам необходимо выбрать из 4 методик обучения ту, которая позволяет выпускникам ПТУ добиваться максимальной производительности. Попытайтесь решить эту проблему с помощью однофакторного дисперсионного анализа.

Рассмотрите также усложненный вариант этой задачи о выборе методики обучения, если стоимости обучения одного учащегося по разным методикам следующие:

Методика 1 — 900 р.;

Методика 2 — 1200 р.;

Методика 3 — 1000 р.;

Методика 4 — 1100 р.

Уровень значимости α взять равным 5%.

2 [27, с. 81]. Требуется выяснить влияние вида галоидного алкила (фактор А) на процесс радикальной полимеризации. Изучалось влияние на выход полимера (Y, %) пяти различных галогеналкилов: $CH_3I(a_1)$, $C_3H_7I(a_2)$, $C_4H_9I(a_3)$, $C_2H_5Br(a_4)$, $C_3H_7Br(a_5)$. Результаты эксперимента с различными галоидными алкилами (фиксированные уровни фактора А) приведены в табл. 36.

Таблица 36

Номер наблюдения	Уровни фактора А				
	a_1	a_2	a_3	a_4	a_5
1	79,80	87,30	42,45	76,00	70,70
2	86,30	69,60	64,30	85,30	64,65
3	86,50	81,75	78,90	72,80	38,50
4	92,30	77,95	61,00	89,00	77,00
5	76,50	83,65	31,30	76,50	91,50
6	87,05	64,80	72,85	87,45	68,00
7	82,50	67,30	58,65	74,50	38,05
8	90,00	75,45	52,50	93,15	79,95

Проведите однофакторный дисперсионный анализ и с помощью линейных ортогональных контрастов сравните между собой разные режимы полимеризации. Значение уровня значимости $\alpha=5\%$.

3. Катализатор для химической реакции получали четырьмя различными способами, которые соответствовали $p=4$ уровням фактора А. В экспериментах измерялась активность катализатора, причем для каждого уровня было сделано по одинаковому числу $n_j=5$ наблюдений. Необходимо проверить равенство средних значений активности, т. е. независимость качества катализатора от способа получения. Результаты измерений приведены в табл. 37.

Таблица 37

Номер наблюдения	Уровни фактора А			
	1	2	3	4
1	56	64	45	42
2	55	61	46	39
3	62	50	45	45
4	59	55	39	43
5	60	56	43	41

Требуется также в случае отклонения нулевой гипотезы использовать метод линейных ортогональных контрастов и сравнить средние попарно. Уровень значимости α взять равным 1%.

4. При разработке технологии получения биологически активного вещества исследовали влияние типа кристаллизатора (фактор А) и типа растворителя (фактор В) на процент выхода годного продукта. При использовании двух различных кристаллизаторов и трех видов растворителя было проведено по пять процессов для каждого сочетания факторов А и В. Результаты представлены в табл. 38. Провести двухфакторный дисперсионный анализ при $\alpha=5\%$.

Таблица 38

Значение фактора А	Значения фактора В		
	1	2	3
1	55; 70; 60; 65; 55	55; 65; 55; 60; 60	55; 58; 60; 65; 60
2	65; 75; 80; 55; 60	55; 70; 65; 70; 60	55; 75; 68; 70; 64

5 [40, с. 218]. Экспериментально проверяется влияние на износоустойчивость детали таких факторов, как материал (фактор А — два вида) и технология изготовления (фактор В — три метода). Данные эксперимента (число месяцев работы детали) приведены в табл. 39.

Таблица 39

Материал (фактор А)	Технология изготовления (фактор В)		
	1	2	3
1	10; 8; 7; 10	8; 12; 14; 12	15; 8; 10; 10
2	12; 8; 8; 7	12; 13; 11; 14	13; 15; 12; 10

Провести двухфакторный дисперсионный анализ. Уровень значимости принять равным 5%.

6 [16, с. 203]. В четырех различных лабораториях производились испытания трех типов опреснителей морской воды. Каждое конкретное испытание дублировалось три раза, результаты испытаний выражены в % остаточной засоленности воды. Результаты испытаний приведены в табл. 40.

Таблица 40

Лаборатория (фактор А)	Тип опреснителя (фактор В)		
	1	2	3
1-я	3,6; 3,8; 4,1	2,9; 3,1; 3,0	2,7; 2,5; 2,9
2-я	4,2; 4,0; 4,1	3,3; 2,9; 3,2	3,7; 3,5; 3,6
3-я	3,8; 3,5; 3,6	3,6; 3,7; 3,5	3,2; 3,0; 3,4
4-я	3,4; 3,2; 3,2	3,4; 3,6; 3,5	3,6; 3,8; 3,7

Применив двухфакторный дисперсионный анализ, установить, какие факторы или их взаимодействие влияют на результаты испытаний. Значение уровня значимости взять равным 5%.

РЕШЕНИЯ

1. Для решения задачи используем программу П 34/34. Вводим исходные данные: 23 П9 4 П8 3,13 П7. Выполняя инструкцию к П 34/34, находим, что групповые средние по четырем методикам обучения равны: $\bar{Y}_{.1}=75$; $\bar{Y}_{.2}=78$; $\bar{Y}_{.3}=71$; $\bar{Y}_{.4}=90$. Далее определяем

$\hat{\eta}^2 = 3{,}5492354 \cdot 10^{-1}$, $F = 3{,}4846242$, $\eta^2_{max} = 6{,}7372712 \cdot 10^{-1}$; $\eta^2_{min} = 3{,}61196 \cdot 10^{-2}$. Поскольку $3{,}4846242 = F_{расч} > F_{0{,}05;\,3;\,19} = 3{,}13$, то нулевую гипотезу о равенстве всех четырех средних между собой отклоняем с $P_{дов} = 95\%$.

Далее, используя программу П 35/34, решаем задачу линейных контрастов. Находим доверительные интервалы для $\Delta\mu_{ij}$:

$$-18{,}973534 \leqslant \Delta\mu_{12} \leqslant 12{,}973534,$$

$$-12{,}57652 \leqslant \Delta\mu_{13} \leqslant 20{,}57652,$$

$$-33{,}533113 \leqslant \Delta\mu_{14} \leqslant 3{,}533113,$$

$$-8{,}973534 \leqslant \Delta\mu_{23} \leqslant 22{,}973534,$$

$$-29{,}995807 \leqslant \Delta\mu_{24} \leqslant 5{,}995807,$$

$$-37{,}533113 \leqslant \Delta\mu_{34} \leqslant -4{,}66887 \cdot 10^{-1}.$$

Анализ доверительных интервалов позволяет сделать заключение о том, что первые три методики дают в среднем (достоверно с $P_{дов} = 95\%$) одинаковые результаты. Не отличаются между собой по результативности обучения и методики 1—4, 2—4. Достоверная разница наблюдается только между методиками 3—4. Казалось бы, что можно выбрать методику 4, но, учитывая, что она дает в среднем такие же результаты, что и первые две методики, целесообразно отдать предпочтение методике 1, т. к. ее стоимость минимальная.

2. Для решения задачи используем программу П 34/34. Вводим исходные данные задачи: 40 П9 5 П8 2,65 П7 ($F_{0{,}05;\,4;\,35} = 2{,}65$). Выполняя инструкцию к П 34/34, находим средние по режимам: $\bar{Y}_{\cdot 1} = 85{,}11875$; $\bar{Y}_{\cdot 2} = 75{,}975$; $\bar{Y}_{\cdot 3} = 57{,}74375$; $\bar{Y}_{\cdot 4} = 81{,}6125$; $\bar{Y}_{\cdot 5} = 66{,}04375$. Далее определяем $\hat{\eta}^2 = 4{,}3568832 \cdot 10^{-1}$, $F = 6{,}755616$; $\eta^2_{max} = 6{,}0659415 \cdot 10^{-1}$, $\eta^2_{min} = 2{,}6478249 \cdot 10^{-1}$. Поскольку $F > F_{табл}$, то делаем вывод о влиянии вида галогеналкила на выход полимера. Далее находим доверительные интервалы для разностей $\Delta\mu_{ij}$:

$$-10{,}870556 \leqslant \Delta\mu_{12} \leqslant 29{,}158056,$$

$$7{,}360694 \leqslant \Delta\mu_{13} \leqslant 47{,}389306,$$

$$-16{,}508056 \leqslant \Delta\mu_{14} \leqslant 23{,}520556,$$

$-0{,}939306 \leqslant \Delta\mu_{15} \leqslant 39{,}089306,$

$-1{,}783056 \leqslant \Delta\mu_{23} \leqslant 38{,}245556,$

$-25{,}651806 \leqslant \Delta\mu_{24} \leqslant 14{,}376806,$

$-10{,}083056 \leqslant \Delta\mu_{25} \leqslant 29{,}945556,$

$-43{,}883056 \leqslant \Delta\mu_{34} \leqslant -3{,}854444,$

$-28{,}314306 \leqslant \Delta\mu_{35} \leqslant 11{,}714306,$

$-4{,}445556 \leqslant \Delta\mu_{45} \leqslant 35{,}583056.$

Из этих интервалов следует, что достоверное различие средних наблюдается между первым и третьим, третьим и четвертым режимами.

3. Введем исходные данные задачи согласно инструкции к П 34/34: 20 П9 4 П8 5,29 П7 ($F_{0,01;\ 3;\ 16}=5{,}29$). Вводя последовательно столбцы, находим средние групповые по режимам: $\bar{Y}_{\cdot 1}=58{,}4$; $\bar{Y}_{\cdot 2}=57{,}2$; $\bar{Y}_{\cdot 3}=43{,}6$; $\bar{Y}_{\cdot 4}=42{,}0$. Продолжая вычисление, определяем $\hat{\eta}^2=8{,}4815498\cdot 10^{-1}$; $F=29{,}790199$; $\eta^2_{max}=9{,}9876624\cdot 10^{-1}$; $\eta^2_{min}=6{,}9754372\cdot 10^{-1}$.

Поскольку $F \approx 29{,}8 > 5{,}29 = F_{табл}$, то у нас есть все основания отклонить нулевую гипотезу о независимости средней активности катализатора от способа получения. Далее, используя программу П 35/34, сравниваем попарно средние активности. При этом получаем следующие доверительные границы для разностей средних $\Delta\mu_{ij}$:

$-7{,}7788195 \leqslant \Delta\mu_{12} \leqslant 10{,}17882,$

$5{,}821181 \leqslant \Delta\mu_{13} \leqslant 23{,}77882,$

$7{,}421181 \leqslant \Delta\mu_{14} \leqslant 25{,}37882,$

$4{,}621181 \leqslant \Delta\mu_{23} \leqslant 22{,}57882,$

$6{,}221181 \leqslant \Delta\mu_{24} \leqslant 24{,}17882,$

$-10{,}57882 \leqslant \Delta\mu_{34} \leqslant 7{,}3788195.$

На основании этих результатов можно сделать вывод о том, что нет значимых, достоверных различий в активности катализатора при получении его способами 1 и 2, 3 и 4. Все остальные попарные сравнения свидетельствуют о значимом различии средних. Поэтому, исходя из очевидной задачи увеличения активности ката-

лизатора, целесообразно детальнее исследовать режимы 1 и 2 (какой из них более дешевый, меньше расход драгметаллов и т. п.) и выбрать окончательно один из них.

4. Вводим в ПМК программу П 38/34—1 и согласно п. 1 инструкции к этой программе по адресу 40 вместо символа В нажимаем клавишу 3, т. к. фактор В имеет 3 уровня, соответственно по адресу 90 вместо символа А набираем число 2, т. к. фактор А имеет 2 уровня (2 типа кристаллизатора). Далее выполняем п. 2 инструкции, вводя вспомогательное число 7 при обработке каждой строки и данные табл. 35.

7 П7 55 В/О С/П (3025) 70 В/О С/П (7925)
60 В/О С/П (11525) 65 В/О С/П (15750) 55 В/О С/П (18775) 5 С/П (3721) 5 С/П (0)
55 В/О С/П (21800) 65 В/О С/П (26025) 55 В/О С/П (29050) 60 В/О С/П (32650) 60 В/О С/П (36250) 5 С/П (7202) 5 С/П (0)
55 В/О С/П (39275) 58 В/О С/П (42639) 60 В/О С/П (46239) 65 В/О С/П (50464) 60 В/О С/П (54064) 5 С/П (10754,16) 5 С/П (0) С/П (0)

Приступаем к вводу второй строки:

7 П7 65 В/О С/П (58289) 75 В/О С/П (63914)
80 В/О С/П (70314) 55 В/О С/П (73339) 60 В/О С/П (76939) 5 С/П (15243,16) 5 С/П (0)
55 В/О С/П (79964) 70 В/О С/П (84864) 65 В/О С/П (89089) 70 В/О С/П (93989) 60 В/О С/П (97589) 5 С/П (19339,16) 5 С/П (0)
55 В/О С/П (100614) 75 В/О С/П (106239)
68 В/О С/П (110863) 70 В/О С/П (115763)
64 В/О С/П (119859) 5 С/П (23748,12) 5 С/П (0) С/П (0).

Ввод двух строк и их раздельная обработка окончена. Выполняем п. 4 инструкции: С/П (377). Выполняем п. 5 инструкции: 30 ПД 2 ПА 3 ПВ 6 ПС. Выполняем п. 6: С/П (118740,6). Прежде чем вводить программу П 38/34—2, выпишем содержимое всех регистров памяти, включая стековые регистры X, Y, Z и T. На регистре X число 118740,6. Вращаем трижды стек и записываем содержимое регистров: F↓ (PY=119859) F↓ (PZ=299,77) F↓ (PT=6). Далее вызываем содержимое регистров прямого доступа: ИП 0 (23748,12) ИП 1 (1885)

и т. д. Ниже приведено содержимое всех этих регистров.

P0=23748,12	P5=118740,6	PA=2
P1=1885	P6=119859	PB=3
P2=7913,6577	P7=299,77	PC=6
P3=9741,69	P8=7913,6577	PД=30
P4=197,4	P9=197,4	

Для чего целесообразно записать отдельно содержимое всех этих регистров памяти? Если, работая с программой П 38/34—2, обнаружите, что нарушили последовательность действий, например, по п. 9 ввели число 100 в регистр 1 вместо регистра 8 и т. п., либо получили результаты, которые свидетельствуют о том, что при наборе программы П 38/34—2 была допущена ошибка, не исправленная впоследствии при проверке программы, можно будет начать счет по второй программе вновь, не выполняя повторно п. 2—6 инструкции. Для этого достаточно будет только ввести числа, выписанные по окончании работы первой программы в соответствующие регистры памяти, и вновь начать счет.

Заметим, что содержимое регистров стека и некоторых регистров прямого доступа совпадает, например PX и P5, PY и P6, PZ и P7, PT и PC. Используя это совпадение, восстанавливаем содержимое регистров стека: ИПС (6) ИП7 (299,77) ИП6 (119859) ИП5 (118740,6). Восстановив содержимое регистров стека, переходим к п. 7 инструкции В/О F ПРГ и начинаем ввод второй программы. По окончании ввода программы и ее проверки переходим в режим работы F АВТ и выполняем промежуточные вычисления: В/О С/П (5,07411). Далее 100 П8 8 П1 С/П ($21{,}13784=\widehat{\eta}^2_x$) С/П ($78{,}862195=\widehat{\eta}^2_z$) С/П ($100=\widehat{\eta}^2_x+\widehat{\eta}^2_z$) С/П ($2{,}8727938\cdot10^{-1}=\widehat{\eta}^2_{AB}$) С/П ($2{,}2326474=\widehat{\eta}^2_B$) С/П ($18{,}617877=\widehat{\eta}^2_A$) С/П ($1{,}2865665=F_x$) С/П ($5{,}6659474=F_A$) С/П ($4{,}3713626\cdot10^{-2}=F_{AB}$) С/П ($3{,}3972892\cdot10^{-1}=F_B$).

Вычисления окончены. Проведем анализ полученных результатов.

Для оценки значимости влияния факторов А, В и их взаимодействия и проверки статистической достоверности полученных оценок $\hat{\eta}^2_i$ сравним полученные расчетные значения F-отношения с табличными значениями критерия Фишера $F_{\alpha;\, f_1;\, f_2}$. Доверительную вероятность возьмем равной 95%. При определении табличных значений критерия Фишера необходимо правильно вычислить степени свободы f_1 и f_2 для каждого критерия.

Для $F_A : f_1 = A-1;\quad f_2 = A \cdot B \cdot (m-1);$

$F_B : f_1 = B-1;\quad f_2 = A \cdot B \cdot (m-1);$

$F_{AB} : f_1 = (A-1) \cdot (B-1);\quad f_2 = A \cdot B \cdot (m-1),$

где А — число уровней фактора А; В — число уровней фактора В; m — число повторных наблюдений в каждой клетке (табл. 35). В нашем случае А=2; В=3; m=5. Из таблицы находим $F_A = 4{,}3$; $F_B = 3{,}4$; $F_{AB} = 3{,}4$. Поскольку только для фактора А $F_{расч} > F_{табл}$, то значимым можно признать только действие этого фактора. Вместе с тем необходимо отметить, что фактор А дает только 18,62% вклада в общую дисперсию. Остальная часть обусловлена неучтенными факторами. Проверка вычислений

$$\hat{\eta}^2_A + \hat{\eta}^2_B + \hat{\eta}^2_{AB} = \hat{\eta}^2_X;\quad \hat{\eta}^2_X + \hat{\eta}^2_Z = 100\%.$$

5. По окончании работы программы П 38/34—1 имеем следующие результаты:

P0=713,4375	P6=2939	PC=6
P1=259	P7=58,7084	PД=24
P2=233,34027	P8=233,34027	PX=2853,75
P3=355,70312	P9=17,5	PY=2939
P4=64,75	PA=2	PZ=58,7084
P5=2853,75	PB=3	P1=6

Вводим программу П 38/34—2 и выполняем п. 8 : В/О С/П (3,0832594). Далее 100 П8 8 П1 С/П (40,781503 $= \hat{\eta}^2_X$) С/П ($\hat{\eta}^2_Z$=59,218496%) С/П ($\hat{\eta}^2_X + \hat{\eta}^2_Z$=100%) С/П ($\hat{\eta}^2_{AB}$=2,141771%) С/П ($\hat{\eta}^2_B$=35,137411%)

С/П ($\hat{\eta}^2_A$=3,5023212%) С/П (F_x=2,4791817)
С/П (F_A=1,0645623) С/П (F_{AB}=3,2550539·10^{-1})
С/П (F_B=5,3401675).

Сравнение полученных F-отношений с табличными позволяет отклонить нулевую гипотезу об отсутствии влияния фактора В, т. к. $F_B=5{,}34>F_{0{,}05;\,2;\,18}=3{,}55$. В то же время принимаются нулевые гипотезы об отсутствии влияния фактора А и эффекта взаимодействия факторов А и В. Таким образом, на износоустойчивость детали оказывает влияние только технология ее изготовления.

6. При вводе программы П 38/34 по адресу 40 набираем число 3 (число уровней фактора В), а по адресу 90 набираем число 4 (число уровней фактора А).

Согласно инструкции к программе П 38/34 вводим данные. После обработки всех строк таблицы по п. 4 инструкции имеем на индикаторе следующий результат: 41,266667. Выполняем п. 5 и 6: 36 ПД 4 ПА 3 ПВ 12 ПС С/П. Результат 430,75332. Вращая операционный стек, выводим на индикатор содержимое его регистров: F↓ (431,32=PY) F↓ (5,01888=PZ) F↓ (12=PT). Извлекаем содержимое регистров прямого доступа: ИП0 (143,58444) ИП1 (123,8) ИП2 (47,405677) ИП3 (35,586806) ИП4 (41,266667) ИП5 (430,75332) ИП6 (431,32) ИП7 (5,01888) ИП8 (47,405677) ИП9 (14,833334) ИПА (4) ИПВ (3) ИПС (12) ИПД (36). Перед набором второй программы П 38/34—2 восстанавливаем содержимое регистров операционного стека: ИПС (12) ИП 7 (5,01888) ИП6 (431,32) ИП5 (430,75332). Переходим к началу программы В/О и переходим в режим программирования F ПРГ. Вводим вторую программу, переходим в режим автоматической работы F АВТ и к началу программы В/О. Выполним п. 8 инструкции: С/П (2,7950469). Выполняем последующие пункты: 100 П8 8 П1 С/П (89,854553=$\hat{\eta}^2_x$) С/П (10,145446=$\hat{\eta}^2_z$) С/П (100=$\hat{\eta}^2_x+\hat{\eta}^2_z$) С/П (50,040585=$\hat{\eta}^2_{AB}$) С/П (23,403186=$\hat{\eta}^2_B$) С/П (16,410782=$\hat{\eta}^2_A$) С/П

($19{,}323575 = F_x$) С/П ($12{,}940412 = F_A$) С/П ($19{,}729279 = F_{AB}$) С/П ($27{,}68121 = F_B$). Начинаем проверку значимости полученных оценок коэффициентов детерминации сравнением соответствующих F-отношений с табличными значениями F-критерия согласно п. 16—19 инструкции. $F_{0{,}05;\ 13;\ 24} = 2{,}18$; $F_{0{,}05;\ 3;\ 24} = 3{,}01$; $F_{0{,}05;\ 6;\ 24} = 2{,}51$; $F_{0{,}05;\ 2;\ 24} = 3{,}4$. Сравнивая эти табличные значения F-критерия с расчетными F-отношениями, делаем вывод о значимости всех полученных оценок коэффициентов детерминации.

ГЛАВА 7

АНАЛИЗ СВЯЗИ МЕЖДУ КОЛИЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ (КОРРЕЛЯЦИОННЫЙ АНАЛИЗ)

Знанию всегда предшествует предположение.

В. Гумбольдт

7.1. Задачи корреляционного и регрессионного анализа

Корреляция в широком смысле этого понятия означает зависимость, соотношение, связь между имеющимися процессами, явлениями, отдельными характеристиками одного или нескольких процессов. Для правильного понимания механизма этой связи и ее использования в нужных целях большое значение имеет правильный выбор вида и формы этой связи. При рассмотрении дисперсионного анализа мы установили, что связи между параметрами процессов могут отличаться по силе ее проявления и влияния.

Понятия о корреляции и регрессии тесно связаны друг с другом. С помощью корреляционного анализа оценивается сила стохастической связи, а с помощью регрессионного анализа оценивается и исследуется ее вид, форма. Оценка значимости результатов обоих видов анализа позволяет проверить их достоверность.

«Среди известных методов прикладной математической статистики регрессионный анализ занимает исключительное положение. Это связано не только с тем, что регрессионный анализ является одним из самых распространенных методов обработки результатов наблюдений, но и с тем, что он, по существу, служит основой целого ряда других разделов математической статистики и прежде всего планирования экспериментов, дисперсионного анализа, многомерного статистического анализа. Регрессионный анализ столь замечательный метод, что он даже воспет в стихах» [24]. С 1806 г., когда известный французский математик А. М. Лежандр изложил метод наименьших квадратов, по сегодняшний день развитие регрессионного анализа продолжается. Основной вклад в развитие корреляционного и регрессионного ана-

лиза внесли такие ученые, как А. М. Лежандр, К. Ф. Гаусс, Ф. Гальтон, К. Пирсон, А. Бравэ, А. А. Чупров и многие другие.

Имеются различные виды корреляции и регрессии. Рассмотрим кратко основные.

1. По характеру проявления связи различают следующие виды корреляции:

а) положительная корреляция: увеличение или уменьшение одной переменной приводит соответственно к увеличению или уменьшению другой переменной;

б) отрицательная корреляция: увеличение одной переменной приводит к уменьшению другой переменной, и наоборот.

2. По числу взаимосвязанных переменных:

а) парная, или простая, корреляция: имеется связь между двумя переменными X и Y;

б) множественная корреляция: имеется связь между переменными, число которых более двух;

в) частная корреляция: связь между двумя переменными при постоянных, зафиксированных значениях других, взаимосвязанных с этими двумя переменными.

3. По форме связи различают следующие виды:

а) линейная корреляция: между взаимосвязанными переменными существуют линейные соотношения;

б) нелинейная корреляция: связь между переменными выражается нелинейными (относительно самих переменных) соотношениями.

4. По типу соединения причины и следствия различают следующие виды корреляции:

а) непосредственная корреляция: исследуемые явления непосредственно, без промежуточных связей, связаны между собой;

б) косвенная корреляция: непосредственной причинно-следственной связи между переменными объективно не существует, связь обусловлена общей для этих переменных причиной;

в) ложная корреляция (нонсенс-корреляция) : имеется чисто формальная связь между явлениями, не имеющая никакого логического объяснения, проявляется лишь в силу количественного соотношения между переменными. Хрестоматийные примеры ложной корреляции: тесная положительная корреляция между количеством импортируемых апельсинов и числом смертных

случаев от онкологических заболеваний за последние 50 лет [23]; положительная корреляция между числом аистов, свивших гнезда в южных районах Швеции, и рождаемостью в эти годы в Швеции [23].

Аналогичную классификацию можно провести и для регрессии.

Рассмотрим основные задачи, решаемые методами корреляционного и регрессионного анализа.

ЗАДАЧИ КОРРЕЛЯЦИОННОГО АНАЛИЗА:

1. Количественное измерение силы, интенсивности связи двух и более явлений, переменных.

2. Отбор и ранжирование факторов по силе связи с исследуемым выходным параметром качества.

3. Обнаружение ранее неизвестных причинно-следственных связей между исследуемыми переменными.

ЗАДАЧИ РЕГРЕССИОННОГО АНАЛИЗА:

1. Идентификация форм зависимости между исследуемыми переменными.

2. Определение, оценка функций регрессии.

3. Оценка неизвестных значений (прогноз) зависимой переменной по известным независимым (входным) переменным с помощью функции регрессии.

В данной главе будут рассмотрены методы корреляционного анализа. Методы регрессионного анализа изложены в последующих главах.

До выполнения расчетов на ПМК по определению тех или иных величин целесообразно представить графически имеющуюся выборку наблюдений в виде двухмерной диаграммы. Каждая точка такой диаграммы имеет координаты X_i, Y_i, соответствующие значениям переменных X и Y в i-том наблюдении. Такую совокупность точек в системе координат X, Y называют корреляционным полем. На рис. 10 приведены корреляционные поля для положительной корреляции (а), отрицательной корреляции (б) и при отсутствии корреляции (в). Изображенные эллипсы называют корреляционными эллипсами.

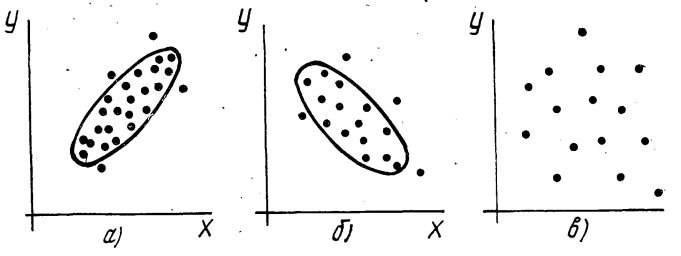

Рис. 10. Примеры корреляционных полей

При обработке двухмерной выборки методами регрессионного анализа определяют параметры прямого уравнения регрессии

$$Y = a_{Y/X} + b_{Y/X} \cdot X, \qquad (7.1.1)$$

и параметры обратного уравнения регрессии (рис. 11):

$$X = a_{X/Y} + b_{X/Y} \cdot Y. \qquad (7.1.2)$$

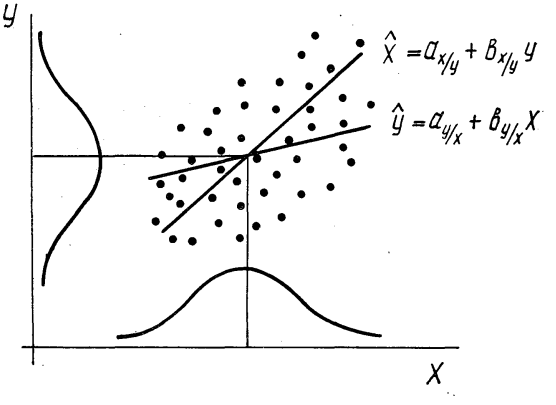

Рис. 11. Соотношение линий прямой и обратной регрессии

Рассмотрим одно из этих уравнений, например (7.1.1). Данное уравнение отражает зависимость не отдельных значений Y_i от X_i, а зависимость значения $\bar{Y}_i$ от X_i (рис. 12).

Связь $\bar{Y}_i$ и Y_i выражается в том, что получаемое по (7.1.1) значение $\bar{Y}_i$ является наиболее вероятным значением величины Y_i при данном X_i. Именно это свойство и используется для интерполяции и экстраполяции значений Y_i (по 7.1.1).

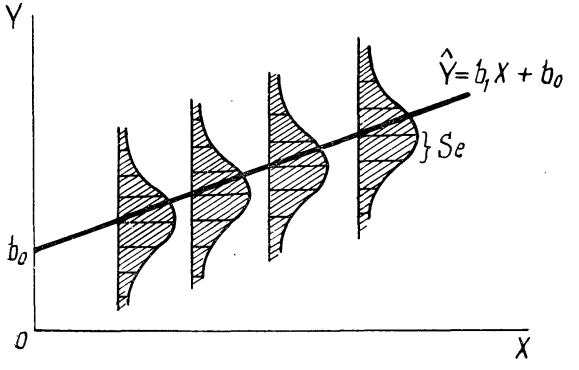

Рис. 12. Соотношение линии регрессии и распределения вероятностей

7.2. Линейный корреляционный анализ

Поскольку при проведении статистического анализа экспериментальных наблюдений мы располагаем только выборкой, то установить истинный вид регрессионной зависимости невозможно. По имеющимся данным мы можем только оценить параметры этой зависимости, рассчитав при этом также и точность полученных оценок. Наиболее часто в регрессионном анализе используют для таких оценок метод наименьших квадратов (МНК), разработку основ которого провели К. Гаусс, Лежандр и П. Лаплас.

Согласно МНК получаемые оценки уравнения регрессии дают такие оценки $\hat{Y}_i$, которые по сравнению с другими оценками $\hat{Y}_i^*$ удовлетворяют следующему требованию:

$$\sigma^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \to \min , \qquad (7.2.1)$$

где Y_i — фактическое значение Y, отвечающее значению X_i.

Рассмотрим последовательность вычисления показателя силы, интенсивности линейной стохастической связи — коэффициента корреляции, который был введен в практику статистического анализа Ф. Гальтоном и К. Пирсоном.

По имеющейся выборке двухмерных наблюдений находим средние значения $\bar{X}$ и $\bar{Y}$ и среднеквадратические отклонения s_X и s_Y. Для получения безразмерной величины силы связи значения X_i и Y_i приведем к стандартизованному виду по формулам:

$$\tilde{X}_i = \frac{X_i - \bar{X}}{s_X}; \quad \tilde{Y}_i = \frac{Y_i - \bar{Y}}{s_Y}.$$

Просуммируем произведения $\tilde{X}_i\tilde{Y}_i$:

$$\sum_{i=1}^{n} \tilde{X}_i\tilde{Y}_i = \sum_{i=1}^{n} \frac{X_i - \bar{X}}{s_X} \cdot \frac{Y_i - \bar{Y}}{s_Y}. \qquad (7.2.2)$$

При большей силе связи переменных X и Y эти суммы также будут больше. Выборочный коэффициент корреляции r_{XY} представляет собой среднее значение такого произведения $\tilde{X}_i\tilde{Y}_i$:

$$r_{XY} = \frac{\sum_{i=1}^{n} \frac{X_i - \bar{X}}{s_X} \cdot \frac{Y_i - \bar{Y}}{s_Y}}{n - 1} = \hat{\rho}_{XY}. \qquad (7.2.3)$$

После несложных преобразований получаем выражение, используемое для практических вычислений:

$$r_{XY} = \frac{n \sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{\sqrt{\left(n \sum_{i=1}^{n} X^2_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i\right) \cdot \left(n \sum_{i=1}^{n} Y^2_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} Y_i\right)}}. \qquad (7.2.4)$$

Величина коэффициента корреляции принимает значения в интервале

$$-1 \leqslant r_{XY} \leqslant +1.$$

Чем ближе r_{XY} к ± 1, тем теснее, интенсивнее связь.

Чем ближе r_{XY} к 0, тем слабее исследуемая связь. Однако если между переменными существует нелинейная связь, то r_{XY} также может быть близок или даже равен нулю. Необходимо помнить, что r_{XY} — это показатель силы линейной связи.

Для вычисления выборочного значения коэффициента корреляции может быть использована приводимая ниже программа П 40/34 [10].

ЗАДАЧИ К ГЛ. 7

1 [58, с. 16]. На основании данных по 20 совхозам Латвии о количестве вносимых в почву минеральных удобрений и урожайности зерновых на этих почвах определить коэффициент корреляции между этими величинами (табл. 41).

Таблица 41

Кол-во удобрений, ц/га	Урожайность, ц/га	Кол-во удобрений, ц/га	Урожайность, ц/га	Кол-во удобрений, ц/га	Урожайность, ц/га	Кол-во удобрений, ц/га	Урожайность, ц/га
1,0	24	0,8	18	1,9	26	2,7	32
2,8	42	1,6	32	1,7	36	2,1	24
1,4	28	1,6	26	2,1	40	1,5	18
1,3	20	1,9	18	2,3	38	2,0	30
2,0	34	1,7	22	2,6	36	2,3	32

Для полученной точечной оценки r построить 95%-ный доверительный интервал и проверить r на значимость при $\alpha = 5\%$.

2 [22, с. 208]. Имеются данные о весе и росте десяти мужчин и восьми женщин. Требуется проверить гипотезу о том, что коэффициент корреляции между ростом и весом для мужчин и женщин совпадает, $\alpha = 5\%$ (табл. 42).

Таблица 42

Мужчины				Женщины			
Рост, см	Вес, кг	Рост, см	Вес, кг	Рост, см	Вес, кг	Рост, см	Вес, кг
166	63	188	85	161	60	170	68
187	98	175	84	158	61	167	69
170	82	183	79	165	68	177	80
171	80	179	75	154	52		
182	86	178	76	166	67		

РЕШЕНИЯ

1. После ввода программы П 40/34 и перевода ПМК в режим вычислений вводим число наблюдений: 20 В/О С/П (0). Далее начинаем ввод наблюдений: 1 С/П (1) 24 С/П (24) 2,8 С/П (2,8) 42 С/П (42) и т. д. ... 2,3 С/П (2,3) 32 С/П ($6{,}8553921 \cdot 10^{-1} = r$). Таким образом, точечная оценка значения генерального коэффициента корреляции между количеством вносимых в почву минеральных удобрений и урожайностью зерновых на этой почве равна $r \approx 0{,}686$.

Для построения 95%-ного доверительного интервала для коэффициента корреляции используем программу П 41/34, при этом получаем $\rho_{max} = 8{,}652297 \cdot 10^{-1}$ и $\rho_{min} = 3{,}4978387 \cdot 10^{-1}$. Поскольку полученный доверительный интервал не покрывает нулевое значение, то с $P_{дов} = 95\%$ можно отклонить $H_0 : \rho = 0$.

2. Формулируем нулевую гипотезу $H_0 : \rho_м = \rho_ж$ и альтернативную гипотезу $H_1 : \rho_м \neq \rho_ж$. Используя программу П 40/34, находим выборочные оценки $r_м = 6{,}7443023 \cdot 10^{-1}$ и $r_ж = 9{,}665771 \cdot 10^{-1}$.

Далее, применяя программу П 42/34, производим сравнение двух коэффициентов корреляции. В результате проверки получаем значение —1, т. е. нулевая гипотеза H_0 отклоняется, что означает разную силу связи роста и веса у мужчин и у женщин.

ГЛАВА 8

ИССЛЕДОВАНИЕ ВИДА ЗАВИСИМОСТИ МЕЖДУ КОЛИЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ (ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ)

В прямой черте от точки к точке...

Ю. Тувим

При анализе имеющихся выборок кроме значения выборочного коэффициента корреляции r необходимо бывает оценить и параметры регрессионных уравнений: $Y = a_{Y/X} \cdot X + b_{Y/X}$ — уравнение прямой регрессии и $X = a_{X/Y} \cdot Y + b_{X/Y}$ — уравнение обратной регрессии. Эти уравнения не являются алгебраическими и не могут быть получены одно из другого, т. к. они отражают стохастическую (вероятностную) связь между Y и X.

Не делая полностью вывод оценок параметров a и b уравнения регрессии, приведем конечные выражения:

$$a_{Y/X} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}; \quad a_{X/Y} = \frac{n\sum XY - \sum X \sum Y}{n\sum Y^2 - (\sum Y)^2}; \tag{8.0.1}$$

$$b_{Y/X} = \frac{\sum Y \sum X^2 - \sum XY \sum X}{n\sum X^2 - (\sum X)^2}; \quad b_{X/Y} = \frac{\sum X \sum Y^2 - \sum XY \sum Y}{n\sum Y^2 - (\sum Y)^2}.$$

На рис. 13 показано расположение линии регрессии в зависимости от комбинации значений a и b.

Прямые, построенные по уравнениям прямой и обратной регрессии, пересекаются в «центре тяжести» ($\bar{X}$; $\bar{Y}$) и образуют «ножницы» (см. рис. 11). Чем уже «ножницы», тем ближе стохастическая связь с функциональной. При функциональной связи $r = \pm 1$ и обе прямые сливаются.

8.1. Точность оценки уравнения регрессии

Качество подбора, оценки функции регрессии можно оценить с помощью стандартных ошибок или дисперсий

остатков и дисперсий оценок параметров уравнения регрессии. Точность получаемых по уравнению регрессии оценок $\hat{Y}_i$ во многом определяется рассеянием фактических значений Y_i вокруг линии регрессии в качестве меры рассеяния используют дисперсию относительно линии регрессии. Для ее определения находят сумму квадратов отклонений (ошибок) фактических значений Y_i от расчетных значений $\hat{Y}_i$, т. е.

$$\sum_{i=1}^{n} e^2_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 .$$

Дисперсия остатков, ошибок равна

$$s^2_{Y/X} = \frac{\sum_{i=1}^{n} e^2_i}{n-2} . \qquad (8.1.1)$$

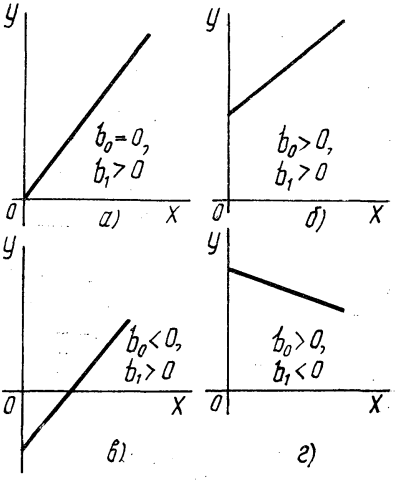

Рис. 13. Расположение линии регрессии в зависимости от комбинации значений параметров a и b

Величина $s^2_{Y/X}$ является выборочной оценкой генеральной дисперсии ошибок ε_i. Величину $s^2_{Y/X}$ можно определить, не рассчитывая значения e_i. Для этого используют имеющиеся соотношения между параметрами корреляции и регрессии, в частности

$$s^2_{Y/X} = \frac{n-1}{n-2} \cdot s_Y^2 (1 - r^2) , \qquad (8.1.2)$$

$$r = \frac{b_{Y/X} \cdot s_X}{s_Y} . \qquad (8.1.3)$$

Очевидно, что не только коэффициент корреляции, но и параметры прямого и обратного уравнений регрессии являются точечными, выборочными оценками соответствующих генеральных параметров. Можно определить стандартные ошибки или дисперсии этих оценок параметров уравнений регрессии. Какова цель определения этих стандартных ошибок параметров регрессии?

Ниже мы покажем, что знание этих ошибок позволяет решать очень важные задачи проверки значимости, достоверности этих оценок параметров регрессии и нахождения доверительных интервалов для линий регрессии. Стандартная ошибка оценки коэффициента регрессии $a_{Y/X}$ определяется по выражению

$$s_{a_{Y/X}} = \frac{s_{Y/X}}{s_X \sqrt{n-1}}, \qquad (8.1.4)$$

где $s_{Y/X}$ — корень квадратный из дисперсии остатков $s^2_{Y/X}$ по (8.1.1) и (8.1.2); s_X — стандартное отклонение переменной X. Из (8.1.4) следует, что $s_{a_{Y/X}}$ зависит

— от рассеяния остатков. Чем больше доля вариации значений переменной Y, необъясненной ее линейной регрессией от X, тем меньше точность оценки $a_{Y/X}$;

— от рассеяния значений независимой переменной X. Чем больше интервал изменения X и рассеяние значений X, тем точнее оценка $a_{Y/X}$, т. к. в выражении (8.1.4) значение s_X стоит в знаменателе. Из этого следует, что при вытянутом вдоль оси X корреляционном поле (эллипсе) мы получаем более надежную оценку углового коэффициента $a_{Y/X}$ уравнения регрессии;

— от объема выборки наблюдений. Чем больше объем наблюдений, тем точнее оценка $a_{Y/X}$.

Теперь, зная оценку углового коэффициента уравнения регрессии $a_{Y/X}$ и точность этой оценки $s_{a_{Y/X}}$, можно проверить статистические гипотезы относительно значения $a_{Y/X}$.

Формулируется нулевая гипотеза $H_0: a_{Y/X} = a_0$, где a_0 — заданное значение, это может быть либо нуль, либо другое определенное отрицательное или положительное число. Когда задается $a_0 = 0$, то такая H_0 означает, по существу, проверку наличия корреляционной зависимости между переменными X и Y. Если проверка H_0 подтвердит, что $a_0 = 0$, то это будет означать, что линия регрессии параллельна оси X и проходит через ось Y в точке $b_{Y/X} = \bar{Y}$... Для проверки H_0 используется t-статистика

$$t = \frac{a_{Y/X} - a_0}{s_{a_{Y/X}}}, \qquad (8.1.5)$$

которая при истинной H_0 подчиняется t-распределению

Стьюдента с числом степеней свободы $f=n-2$. В случае $|t|>t_{\alpha;f}$ гипотеза H_0 отвергается, а при $|t|<t_{\alpha;f}$ H_0 принимается.

Стандартная ошибка свободного члена, константы $b_{Y/X}$ в уравнении регрессии определяется по формуле

$$s_{b_{Y/X}} = s_{Y/X} \cdot \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1) \cdot s^2_X}} . \qquad (8.1.6)$$

Очевидно, что и в этом случае увеличение объема выборки n и дисперсии s^2_x приводит к повышению точности оценки свободного члена $b_{Y/X}$. Зная оценку $b_{Y/X}$ и определив $s_{b_{Y/X}}$ по (8.1.6), можно провести аналогичную проверку значимости, достоверности полученной оценки $b_{Y/X}$. Для проверки $H_0: \beta_{Y/X}=b_0$ используют также аналогичную t-статистику

$$t = \frac{b_{Y/y} - b_0}{s_{b_{Y/X}}} , \qquad (8.1.7)$$

которая при истинной H_0 подчиняется t-распределению Стьюдента с числом степеней свободы $f=n-2$. В случае $|t|>t_{\alpha;f}$ гипотеза H_0 отвергается, а при $|t|<t_{\alpha;f}$ H_0 принимается. При значении $b_0=0$ формулируемая H_0 означает, что мы выдвигаем предположение о том, что линия регрессии проходит через центр координат X и Y (0; 0), т. е. что значение $\hat{Y}_i=0$ при $X_i=0$.

Обратим внимание еще на одну важную связь корреляционных и регрессионных параметров. Обратимся к выражению (8.1.2). При объеме выборки $n\to\infty$ мы получим вместо стоящих в этом выражении значений выборочных оценок $s^2_{Y/X}$, s^2_Y и r^2 соответствующие значения генеральных параметров $\sigma^2_{Y/X}$, σ^2_Y и ρ^2. Учитывая, что при этом отношение $(n-1)/(n-2)\to 1$, вместо (8.1.2) получим следующее выражение:

$$\sigma^2_{Y/X}=\sigma^2=\sigma^2_Y(1-\rho^2). \qquad (8.1.8)$$

Выполнив несложные преобразования, получим:

$$\rho^2=(\sigma^2_Y-\sigma^2)/\sigma^2_Y, \qquad (8.1.9)$$

т. е. квадрат коэффициента корреляции равен доле дисперсии Y, объясненной знанием X. Из (8.1.8) следует, что $\sigma=\sigma_Y\sqrt{1-\rho^2}$, т. е. оставшаяся «необъясненной» че-

рез вариацию X доля стандартного отклонения σ равна $\sqrt{1-\rho^2}$. Так как σ выражается, в тех же единицах измерения, что и Y, то это позволяет более осмысленно интерпретировать смысл ρ. В табл. 43 показано соотношение между коэффициентом корреляции ρ и двумя его наиболее связанными между собой интерпретациями. Отметим, что даже при таких больших значениях $\rho=0{,}9$ и $\rho=0{,}95$ большáя доля стандартного отклонения 0,44 и 0,31 соответственно остается необъясненной через вариацию X.

Таблица 43

Величина коэффициента корреляции	Процент объясненной дисперсии	Процент стандартного отклонения величины, необъясненной вариацией величины X
00	0	100,0
±0,2	4	97,98
±0,4	16	91,65
±0,6	36	80,00
±0,8	64	60,00
±0,9	81	43,59
±0,95	90,25	31,22
±0,99	98,01	14,11

Зная оценки $a_{Y/X}$ и $b_{Y/X}$, а также стандартные отклонения этих оценок, несложно построить доверительные интервалы для этих оценок по формулам:

$$a_{Y/X} \pm t_{\alpha;\,f} \cdot s_{a_{Y/X}}, \qquad (8.1.10)$$

$$b_{Y/X} \pm t_{\alpha;\,f} \cdot s_{b_{Y/X}}, \qquad (8.1.11)$$

где $t_{\alpha;\,f}$ — квантиль t-распределения Стьюдента с $f=n-2$. Полуширину доверительного интервала $t_{\alpha;\,f} \cdot s$ называют **точностью оценки** параметра регрессии. С учетом доверительных интервалов для этих двух параметров линейного уравнения парной регрессии мож-

но построить доверительную зону в осях X—Y для истинной линии регрессии, в уравнение которой входят не оценки a и b, а их генеральные аналоги α и β. Такая зона будет ограничена двумя кривыми линиями-гиперболами, и расположена внутри по отношению к этим линиям (рис. 14). На этом рисунке пунктиром показаны несколько из возможных положений прямой истинного,

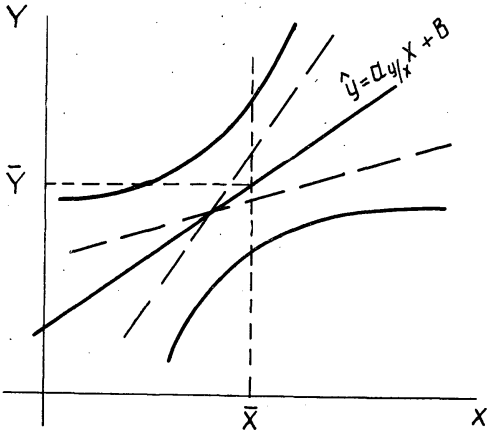

Рис. 14. Доверительная зона для истинной линии регрессии

генерального уравнения регрессии. Из рисунка видно, что наиболее точно по оцененному уравнению регрессии можно предсказать значение $\hat{Y}_i$ при $X_i=\bar{X}$, в этом случае $\hat{Y}_i=\bar{Y}$. Этому значению $X_i=\bar{X}$ отвечает минимум доверительного интервала для $\hat{Y}_i$. По мере удаления X_i от $\bar{X}$ точность прогноза с помощью уравнения регрессии уменьшается, т. е. увеличивается доверительный интервал для оценки $\hat{Y}_i$. Дисперсия таких отдельных значений $\hat{Y}_i$, прогнозируемых с помощью уравнения регрессии и $X=X_i$, определяется формулой

$$s^2_{\hat{Y}_i} = s^2_{Y/X}\left[1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum\limits_{i=1}^{n}(X_i-\bar{X})^2}\right]. \quad (8.1.12)$$

С помощью программ П 43/34 и П 44/34 оценивают не только параметры прямого уравнения регрессии $a_{Y/X}$ и $b_{Y/X}$, но и стандартную ошибку оценки по уравнению регрессии.

8.2. Анализ остатков и последовательностей

Важным этапом корреляционно-регрессионного анализа является анализ остатков, т. е. анализ значений величины $e_i = Y_i - \hat{Y}_i$, где Y_i — фактическое значение переменной Y при значении $X = X_i$, а $\hat{Y}_i$ — оценка Y_i по уравнению регрессии. На рис. 15 представлены три реализации экспериментальных данных равного объема с идентичными гистограммами и равными средними и дисперсиями. Однако они существенно отличаются между собой. На рис. 15,*а* значения измеряемой величины распределены случайно на всем отрезке времени получения выборки. Интервал изменения самой величины равен примерно $\pm 2s$ вдоль всей оси i. На рис. 15, *б* совершенно иное расположение значений переменной X. Видно, что по мере увеличения номера наблюдения i значение X_i в среднем возрастает. В данном случае мы имеем дело с дрейфом, трендом во времени среднего значения мгновенных распределений. На последнем рисунке видно, что среднее значение остается постоянным, но по мере увеличения номера измерения происходит увеличение рассеивания, дисперсии наблюдений. Если в первом случае число наблюдений (длину реализации) увеличить, допустим, в 2—3 раза, то при сохранении имеющейся тенденции это не приведет к сколько-нибудь значительному изменению и среднего, и дисперсии. Если то же самое сделать во втором случае (см. рис. 15, *б*), то существенно изменятся как среднее, так и дисперсия. Наконец, в третьем случае (рис. 15, *в*) может произойти небольшое изменение среднего и значительное увеличение дисперсии.

Если вертикальную ось X заменить на ось e, то можно выделить четыре наиболее часто встречающихся случая взаимного расположения остатков.

1. Остатки распределены нормально и случайно вблизи горизонтальной оси $e = 0$ (см. рис. 15, *а*).

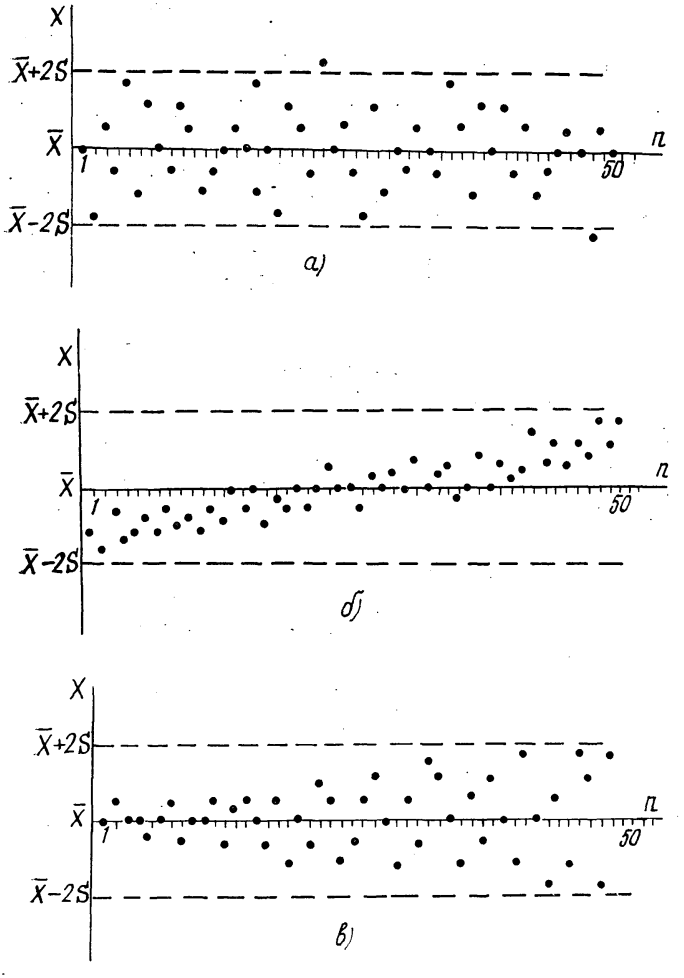

Рис. 15. Три реализации наблюдений

2. По мере увеличения номера наблюдения (измерения) значения остатков увеличиваются и меняют знак «+» на «—», или наоборот (см. рис. 15, *б*).

3. По мере увеличения номера наблюдения (измерения) абсолютное значение, модуль, остатков возрастает либо уменьшается (рис. 15, *в*).

4. По мере увеличения номера наблюдения (измерения) скопление остатков располагается в виде кривой полосы, дважды, трижды пересекающей горизонтальную ось $e=0$.

В первом случае мы имеем адекватную линейную модель с ошибкой e, во втором, третьем и четвертом случаях мы имеем дело с коррелированными остатками. В этом случае линейная модель регрессии неадекватна и необходима более сложная модель, нелинейная. Не рассматривая подробно все три случая, остановимся подробнее на втором. Как правило, такие коррелированные остатки могут наблюдаться при анализе нелинейных зависимостей и временнóго ряда, т. е. выборки, включающей в себя наблюдения за одной и той же величиной в различные, как правило, периодические моменты времени. Для проверки коррелированности остатков часто используют критерий сериальной корреляции Дурбина—Ватсона [55], определяемый выражением

$$D = \frac{\sum_{t=1}^{n} (E_t - E_{t-1})^2}{\sum_{t=1}^{n} E_t^2}, \qquad (8.2.1.)$$

где E_t — остаток $(Y_t - \hat{Y}_t)$ в момент времени t, а $(E_t - E_{t-1})$ — последовательная разность. Возможные значения этой статистики лежат в интервале $0 \leqslant D \leqslant 4$. Если остатки не коррелированы между собой, то значения D располагаются вблизи числа 2. При сильной положительной корреляции остатков величина D близка к нулю. Сильная отрицательная корреляция остатков наблюдается при значениях D, приближающихся к 4. В таблице [17] даны значения верхних $D_и$ и нижних D_l границ критерия. Если величина D, вычисленная по (8.2.1), оказывается меньше, чем D_l, или больше, чем $4-D_l$, то следует считать, что корреляция имеется. Если же значение D заключено внутри интервала $D_и$ и $4-D_и$, то можно считать, что корреляция остатков отсутствует.

Наиболее же часто используют для проверки автокорреляции членов ряда коэффициент автокорреляции.

Члены временного ряда взаимосвязаны между собой, т. е. ряд автокоррелирован, если существует корре-

ляция между последовательными уровнями ряда. Такая корреляция между последовательными, или смещенными на период τ, значениями одного и того же ряда наблюдений называется автокорреляцией переменных. Если $\tau=1$, то говорят о коэффициенте автокорреляции первого порядка [23]. Члены ряда при этом смещаются на одну единицу времени. Вычисление коэффициента автокорреляции при этом производят так же, как и обычного коэффициента корреляции. По сути дела, при нахождении коэффициента автокорреляции мы из одного ряда наблюдений образуем 2 ряда. Например, исходный ряд имеет вид

$X_1, \quad X_2, \quad X_3, \quad X_4, \quad X_5, \quad X_6, \ldots, \quad X_n.$

При $\tau=1$ имеем 2 ряда, аналогичных рядам значений X_i, Y_i: $X_1, X_2, X_3, X_4, X_5, X_6, \ldots, X_{n-1}$ (X_i)

$X_2, X_3, X_4, X_5, X_6, X_7, \ldots, X_n$ (Y_i)

При $\tau=2$ эти 2 ряда будут выглядеть следующим образом: $X_1, X_2, X_3, X_4, X_5, \ldots, X_{n-2}$ (X_i)

$X_3, X_4, X_5, X_6, X_7, \ldots, X_n$ (Y_i)

и т. д.

Между статистикой D Дурбина—Ватсона и коэффициентом автокорреляции ρ_τ существует приближенное соотношение

$$D \approx 2(1-\rho_\tau). \qquad (8.2.2)$$

Как и обычный коэффициент корреляции, значение ρ_τ находится в интервале $-1 \leqslant \rho_\tau \leqslant +1$. График зависимости $\rho_\tau = f(\tau)$ называется коррелограммой, а сама зависимость $\rho_\tau = f(\tau)$ называется автокорреляционной функцией. С помощью программы П 45/34 производится оценка автокорреляционной функции [6] $\rho_\tau = f(n\tau)$, где $n=0$; 1; 2; 3; 4; 5. При $\tau=0$ получаем значение $f(0)=1{,}0$, поскольку коэффициент корреляции величины X с самой собой равен по определению 1.

Для проверки независимости и стационарности ряда наблюдений, помимо рассмотренных выше методов, используется много других критериев, например, критерий «восходящих» и «нисходящих» серий и др. [3]. Очень удобен и прост для понимания и применения критерий Аббе — критерий квадратов последовательных разностей [3]. Наиболее продуктивно применение этого критерия при проверке случайного характера выборки из гене-

ральной совокупности при альтернативном предположении о возможном систематическом смещении среднего по мере увеличения объема выборки.

Для проверки стохастической независимости членов ряда последовательных наблюдений подсчитывается величина

$$\gamma(n) = q^2(n)/s^2(n), \qquad (8.2.3)$$

где

$$q^2(n) = \frac{1}{2\cdot(n-1)} \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2, \qquad (8.2.4)$$

а $s^2(n)$ — выборочная оценка дисперсии по выборке объема n:

$$S^2(n) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. \qquad (8.2.5)$$

В том случае, если $\gamma(n) \leqslant \gamma_\alpha^{min}(n)$, (8.2.6)
то гипотеза о стохастической независимости результатов наблюдения отвергается. Для выборок объемом $n > 60$ критическая величина критерия $\gamma_\alpha^{min}(n)$ подсчитывается по формуле

$$\gamma_\alpha^{min}(n) \approx 1 + \frac{z_\alpha}{\sqrt{n + 0{,}5\,(1+z^2_\alpha)}}, \qquad (8.2.7)$$

где z_α — квантиль нормированного нормального распределения. Относительная ошибка при вычислении $\gamma_\alpha^{min}(n)$ по (8.2.7) не превышает 0,1% [17, 18].

С помощью программы П 46/34 производится вычисление значения критерия Аббе $\gamma(n)$.

ЗАДАЧИ К ГЛ. 8

1 [29, с. 317]. В табл. 44 приведены результаты эксперимента по изучению метода для измерения скорости крови. Здесь X — непосредственные измерения скорости крови, Y — скорость крови, оцененная с помощью изучаемого косвенного метода.

На основе имеющихся данных провести корреляционно-регрессионный анализ, проверить значимость выборочного коэффициента корреляции, построить 95%-ные доверительные интервалы для коэффициента уравнения регрессии $a_{Y/X}$, $b_{Y/X}$.

Таблица 44

X	1190	1455	1550	1730	1745	1770	1900	1920	1960	2295
Y	1115	1425	1515	1795	1715	1710	1830	1920	1970	2300
X	2335	2490	2720	2710	2530	2900	2760	3010		
Y	2280	2520	2630	2740	2390	2800	2630	2970		

Таблица 45

Число колец X	1	2	3	4	5	6	7	8	9	10	11
Возраст Y	3,4	4,4	5,5	6,5	7,4	8,3	9,5	10,5	11,5	12,4	13,3

Таблица 46

t	0	4	10	15	21	29	36	51	68
X	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1

Таблица 47

Y, см	0,30	0,10	0,08	0,32	0,40	0,40	0,50	0,68	0,70	0,85
X, °C	—10	—5	—4	—10	—12	—15	—16	—30	—31	—50
Y, см	0,80	0,85	0,80	0,35	0,75	0,80	0,80	0,78	0,80	0,80
X, °C	—45	—65	—70	—75	—40	—45	—45	—44	—42	—60

2 [47, с. 228]. В некоторых случаях имеет смысл определение возраста коров по числу колец на рогах. Связь между числами колец на рогах и возрастом возникает оттого, что каждый отел, происходящий обычно ежегодно, оставляет на рогах коровы кольцо, отражающее замедление роста рога в периоды глубокой стельности, когда главная масса питательных веществ тратится на питание плода. Провести регрессионный анализ данных, приведенных в табл. 45, с помощью программы П 43/34.

3 [2, с. 407]. На цементном заводе в процессе производства ежедневно в течение 45 дней брались пробы и определялось среднее сопротивление сжатию контрольных кубов (н/см^2 или кг/см^2). Результаты наблюдения: 40, 33, 75, 18, 62, 33, 38, 69, 65, 100, 124, 91, 79, 42, 63, 23, 47, 52, 98, 97, 73, 85, 88, 40, 42, 51, 23, 75, 52, 126, 90, 111, 92, 109, 72, 28, 56, 17, 52, 68, 75, 102, 107, 88, 45.

Проверить стохастическую независимость результатов наблюдения с помощью критерия Аббе. Уровень значимости $\alpha=5\%$.

4. В «Основах химии» Д. И. Менделеев приводит следующие данные о количестве (X) азотно-натриевой соли $NaNO_3$, которое можно растворить в 100 г воды в зависимости от температуры (t):

Провести корреляционно-регрессионный анализ этих данных при значении $\alpha=5\%$ (табл. 46).

5 [41, с. 62]. Провести корреляционно-регрессионный анализ зависимости толщины ледяного покрова (Y) от суммы среднесуточной температуры воздуха (X) (табл. 47).

6 [52, с. 134]. Провести корреляционно-регрессионный анализ зависимости между весом 100 семян ячменя (Y) и продолжительностью вегетационного периода ячменя (X). Уровень значимости $\alpha=5\%$ (табл. 48).

Таблица 48

Y, г	47,5	46,75	45,75	42,85	44,76	41,44	37,00
X, дни	90	85	80	75	70	65	60

РЕШЕНИЯ

1. Вводим программу П 44/34, по п. 2 инструкции вводим вспомогательные числа: 2 ПО 1 П6. Начинаем ввод таблицы: 1190↑1115 В/О С/П (1326850) 1455↑1425 В/О С/П и т. д. ... 3010↑2970 В/О С/П (86125825). Переходим к выполнению п. 4—11 инструкции. С/П ($9{,}7456087 \cdot 10^{-1} = a_{Y/X}$) С/П ($15{,}3534 = b_{Y/X}$) С/П ($2826{,}611 = s^2_{Y/X}$). Итак, искомое уравнение регрессии имеет следующий вид:

$$Y = 9{,}7456087 \cdot 10^{-1} \cdot X + 15{,}3534.$$

Аппроксимация опытных данных этим уравнением дает дисперсию $s^2_{Y\ X} = 2826{,}611$, или $s_{Y/X} = 53{,}165882$. Находим коэффициент корреляции: ИП7 ($9{,}9471149 \cdot 10^{-1} = r$) и обратное уравнение регрессии ИПВ ИПА ИП4÷ ($1{,}0152787 = a_{X/Y}$) ИПС×—($7{,}2509 = b_{X/Y}$). Значит обратное уравнение регрессии можно записать так:

$$X = 1{,}0152787 \cdot Y + 7{,}2509.$$

Естественно, что коэффициент корреляции остался тем же самым, т. к. $r_{XY} = r_{YX}$. Вычислим t-статистику для проверки значимости r: ИП7 1 ИП7—F√÷ИП5 2 —F√× ($54{,}71304 = t_{расч}$). По таблице [17, 18] найдем значение $t_{\alpha/2;\,f}$, где $f = 18 - 2 = 16$, а $\alpha/2 = 0{,}025$ (проверяется двухсторонняя гипотеза, т. е. $H_0 : \rho \neq 0$): $t_{табл} = 2{,}12$. Поскольку расчетное значение t-статистики больше табличного, то гипотеза H_0 отклоняется.

Построим 95%-ный доверительный интервал для коэффициента $a_{Y/X}$. Вспомним, что доверительный интервал для $a_{Y/X}$ строится по формуле $a_{Y/X} \pm t_{\frac{\alpha}{2};\,f} \cdot s_{a_{Y/X}}$. Поскольку интервал строим двухсторонний, то значение $t_{табл} = 2{,}12$. Находим стандартное отклонение для коэффициента $a_{Y/X}$: $s_{a_{Y/X}} = \frac{s_{Y/X}}{s_X \sqrt{n-1}}$. Все необходимые для нахождения этого стандарта величины хранятся в регистрах памяти ПМК. Программа нахождения $s_{aY/X}$

будет выглядеть так: ИП2 ИП6 ИП5 1—×÷F√. Выполнив вручную эту программу, получим значение $s_{aY/X}=$ $=2{,}371764\cdot10^{-2}$. Далее находим значение $\Delta_a=t\cdot s_a$, для этого выполняем: 2,12×(5,0281396·10⁻²=Δ_a) П9 ИПО+ +(1,0248423=$a_{Y/X}^{max}$) ИПО ИП9 — (9,2427947·10⁻¹= $=a_{Y/X}^{min}$).

Поскольку доверительный интервал для углового коэффициента покрывает значение $a_{Y/X}=1$, то на основании этого может быть принята нулевая гипотеза $\alpha_{Y/X}=1$.

Определим стандарт $s_{b_{Y/X}}=s_{Y/X}\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{(n-1)\cdot s^2_X}}$.

Программа нахождения этой величины будет выглядеть так: ИПВ FX² ИП6 ИП5 1—×÷ИП5 F 1/X+ИП2×F√. Выполнив вручную эту программу, получим значение $s_{b_{Y/X}}=52{,}855672$. Находим $\Delta_b=t\cdot s_b$: 2,12×(112,05402= $=\Delta_b$) П9. Далее ИПД+(127,40742=$b_{Y/X}^{max}$) ИПД ИП9 — (—96,70062=$b_{Y/X}^{min}$). Таким образом, доверительный интервал для $b_{Y/X}$ покрывает значение $b_{Y/X}=0$, что позволяет принять нулевую гипотезу $H_0:\beta_{Y/X}=0$.

2. Применив программу П 43/34, получим, что искомое уравнение имеет вид:

$$Y=2{,}4490909+9{,}9636363\cdot10^{-1}\cdot X,$$

$$s=8{,}1901023\cdot10^{-2};\ n=11.$$

3. Вводим программу П 46/34, обрабатываем первое наблюдение: 40 П8 В/О С/П (40). Далее вводим все остальные 44 наблюдения: 33 В/О С/П (33) 75 В/О С/П (75)... 77 В/О С/П (77) 45 В/О С/П (45). Введены все 45 наблюдений. Вычисляем значение критерия Аббе $\gamma(n)$: С/П (5,7343115·10⁻¹).

По таблице [17] находим $\gamma_\alpha^{(min)}$ (45), которое равно 0,7603. Поскольку $\gamma_\alpha^{min}(45)>\gamma(45)$, то гипотеза о стохастической независимости наблюдений отвергается с $\alpha=5\%$.

4. Производим обработку по П 44/34: 2 ПО 1 П6. Вводим исходные данные: 0↑66,7 В/О С/П (0) 4↑17 В/О С/П (9489,89)... и т. д. Выполнив все пункты инструкции к П 44/34, получим следующие числа: 8,7064039·10⁻¹; 67,57794; 9,2042857·10⁻¹; 81,755315; 9,9895495·10⁻¹; 26; 90,144444; 507,5; 385,49777;

—77,32179; 1,1461804; 1,2115714. Идентифицировав эти числа, запишите уравнение, выражающее зависимость растворимости $NaNO_3$ от температуры.

5. $a_{Y/X} = -8{,}2896155 \cdot 10^{-3}$; $b_{Y/X} = 2{,}9706073 \cdot 10^{-1}$;
$s^2_{Y/X} = 3{,}4799955 \cdot 10^{-2}$; $s^2_X = 498{,}01052$;
$s^2_Y = 6{,}7190526 \cdot 10^{-2}$; $\bar{X} = -35{,}7$; $\bar{Y} = 5{,}93 \cdot 10^{-1}$;
$r = -7{,}136736 \cdot 10^{-1}$; $a_{X/Y} = -61{,}441932$;
$b_{X/Y} = 7{,}35066 \cdot 10^{-1}$; $t_{расч} = -2{,}3129798$.

6. Приведенные ниже числовые данные идентифицируйте как результаты анализа исходной таблицы по П 44/34: 2,68252; 20,626785; $3{,}0792857 \cdot 10^{-1}$; 75,0; 43,721428; 116,66666; 13,297642; 2,7016066; $9{,}1208654 \cdot 10^{-1}$; —41,1181; 6, 8784923. После идентификации этих результатов сделайте оценку веса 100 семян ячменя (по уравнению регрессии), если продолжительность вегетационного периода составляет 72 дня.

ГЛАВА 9

НЕЛИНЕЙНЫЙ КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Природа прямизны не знает...

Ю. Тувим

9.1. Нелинейная корреляция

9.1.1. Корреляционное отношение и коэффициент детерминации

В том случае, когда расположение точек на корреляционном поле и проверка значимости линейного уравнения регрессии указывают на нелинейный характер связи переменных X и Y, используют методы нелинейного корреляционного-регрессионного анализа. Изучим основные особенности этого анализа, его взаимосвязь с линейным анализом и программы аппроксимации экспериментальных наблюдений нелинейными зависимостями.

Рассмотрим следующий пример. При разработке технологии полимеризации нового химического продукта было проведено 50 процессов. В данных процессах изучалось влияние расхода катализатора на процент выхода полимера. При этом было использовано 10 различных уровней расхода катализатора. При одном и том же расходе катализатора процесс дублировался пять раз. В табл. 49 представлены результаты экспериментов по исследованию зависимости процента выхода полимера (Y) от расхода катализатора (X). Результаты, представленные в третьей колонке этой таблицы, нанесем в виде точек на плоскости X—Y и соединим штриховой линией (рис. 16). Поскольку без результатов специального анализа, который мы проведем ниже, нельзя сделать достоверного утверждения о том, линейная или нелинейная зависимость имеется на самом деле, то на этом рисунке проведена также линия регрессии, определяемая уравнением $Y = a \cdot X + b$.

Таблица 49

Расход катализатора X_i, г	Процент выхода полимера при постоянном значении X_i, %					Условный средний процент выхода полимера $\bar{Y}_i$, %
	Y_{ij}					
11	37	33	34	35	34	34,6
12	33	34	32	33	31	32,6
13	33	31	32	31	30	31,4
14	31	33	32	33	31	32,0
15	35	34	36	33	34	34,4
16	39	41	39	40	40	39,8
17	46	47	45	45	46	45,8
18	57	55	56	57	54	55,8
19	63	67	67	64	68	65,8
20	80	79	81	78	77	79,0

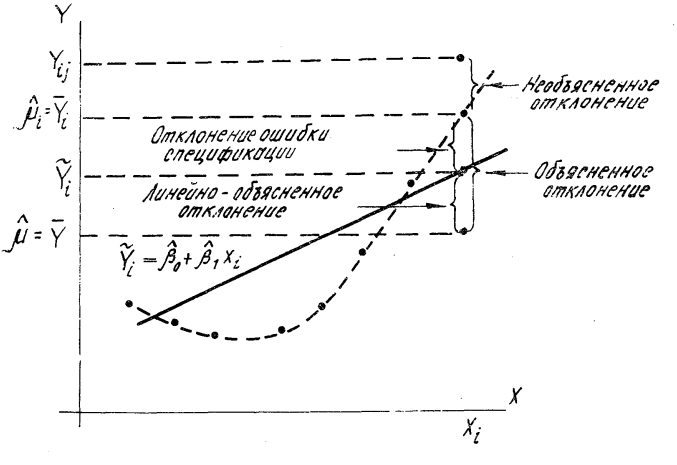

Рис. 16. Линейная и нелинейная спецификация уравнения регрессии

Значения Y_i, определяемые по этому уравнению, обозначим $\hat{Y}_i$. Обозначим через $\bar{Y}_i$ условное среднее, отвечающее данному значению X_i, а через $\bar{Y}_{..}\bar{V}$ общее среднее всех наблюдений. Эти значения являются выборочными оценками соответствующих генеральных средних: условного среднего μ_i и среднего генеральной совокупности $\mu_{..}$. Возьмем отдельную точку (одно наблюдение) на нашем рисунке с координатами $(X_i; Y_{ij})$. Согласно рисунку можно записать следующее равенство:

$$(Y_{ij} - \bar{Y}_{..}) = (\bar{Y}_i - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}), \quad (9.1.1)$$

т. е. общее отклонение = объясненное отклонение + необъясненное отклонение. Аналогичный анализ мы проводим в главе 6 при рассмотрении дисперсионного анализа. Возьмем в квадрат отдельные части выражения (9.1.1):

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{..})^2=\sum_{i=1}n_i(\bar{Y}_i-\bar{Y}_{..})+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2. \quad (9.1.2)$$

Здесь K — число уровней переменной X (в нашем случае число разных значений расхода катализатора); n_i — число наблюдений Y для каждого уровня X_i, т. е. $K = 10$; $n_i = 5$. Выражение (9.1.2) означает: общая вариация = объясненная вариация + необъясненная вариация.

Очевидно, что коэффициент детерминации равен

$$\eta^2 = \frac{\text{объясненная вариация}}{\text{общая вариация}}. \quad (9.1.3)$$

Если наши условные средние $\bar{Y}_i$ не будут зависеть от X_i, то ясно, что пунктирная линия будет проходить параллельно оси X через точку $\bar{Y}_{..}$ на оси Y. С ней сольется и линия регрессии, т. е. в этом случае коэффициент a станет равным нулю, а b станет равным $\bar{Y}_{..}$. Как уже упоминалось выше, корень квадратный из η^2 называют корреляционным отношением: $0 \leqslant \eta \leqslant 1$. При чисто-функциональной связи между X и Y имеем $\eta = 1$. В нашем примере, выполнив суммирование квадратов в (9.1.2), получим $12375{,}28 = 12305{,}28 + 70{,}0$. Тогда $\eta^2 = 12305{,}28/12375{,}28 = 9{,}9434356 \cdot 10^{-1}$, а $\hat{\eta} = 9{,}9716776 \cdot 10^{-1}$.

Проверка значимости корреляционного отношения производится следующим образом. Формулируем нулевую гипотезу $H_0: \eta = 0$, т. е. связь отсутствует, при

$H_1: \eta \neq 0$, т. е. связь существует, где η — корреляционное отношение генеральной совокупности. При истинной нулевой гипотезе F-отношение

$$F_{расч} = \frac{(\text{объясненная вариация})/(K-1)}{(\text{необъясненная вариация})/(N-K)} =$$

$$= \frac{\hat{\eta}^2/(K-1)}{(1-\hat{\eta}^2/(N-K)} \qquad (9.1.4)$$

подчиняется F-распределению Фишера—Снедекора со степенями свободы

$$f_1 = K - 1 \text{ и } f_2 = N - K, \text{ где } N = \sum_{i=1}^{k} n_i.$$

В нашем примере

$$F_{расч} = \frac{12305{,}28/9}{70{,}0/40} = 781{,}2876 \approx 781{,}29.$$

Табличное значение $F_{0,05;\,9:\,40} = 2{,}12$. Поскольку $F_{расч} > F_{0,05;\,9:\,40}$, нулевая гипотеза отвергается и делается вывод о достоверности связи процента выхода полимера с расходом катализатора.

9.1.2. Проверка гипотезы линейности

Теперь, зная, что связь достоверна, необходимо определить форму связи: линейная или нелинейная. Этот этап регрессионного анализа называется спецификацией формы связи. С этой целью объясненное отклонение в (9.1.1) разложим на 2 составляющие. Предположим, что связь линейная. Тогда первое слагаемое правой части равенства (9.1.1) можно записать следующим образом (см. рис. 16):

$$(Y_i - \bar{Y}_{..}) = (\hat{Y}_i - \bar{Y}_{..}) + (\bar{Y}_i - \hat{Y}_i), \qquad (9.1.5)$$

где $\hat{Y}_i$ — есть оценка Y по линейному уравнению при $X = X_i$. Разность $(\bar{Y}_i - \hat{Y}_i)$ назовем ошибкой спецификации, т. к. в первом приближении мы специфицировали зависимость как линейную (сравните с линией регрес-

сии на рис. 16). С учетом ошибки спецификации выражение 9.1.1) примет следующий вид:

$$(Y_{ij} - \bar{Y}_{..}) = (\hat{Y}_i - \bar{Y}_{..}) + (\bar{Y}_i - \hat{Y}_i) + (Y_{ij} - \bar{Y}_i), \quad (9.1.6)$$

что означает: общее отклонение = линейно-объясненное отклонение + отклонение ошибки спецификации + необъясненное отклонение.

Очевидно, что если истинная зависимость $Y = f(X)$ линейна, то отклонение ошибки спецификации будет равно нулю (см. рис. 16) и объясненное отклонение станет равным линейно-объясненному отклонению. Перейдя, как и выше, к суммам квадратов, получим

$$\sum_{i=1}^{k}\sum_{i=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{k} n_i (\hat{Y}_i - \bar{Y}_{..})^2 +$$

$$+ \sum_{i=1}^{k} n_i (\bar{Y}_i - \hat{Y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2, \quad (9.1.7)$$

что озачает: общая вариация = линейно-объясненная вариация + вариация ошибки спецификации + необъясненная вариация. Переходя от вариаций к дисперсиям, мы можем провести с помощью дисперсионного анализа проверку статистических гипотез о форме связи. Формулируем H_0: зависимость линейна, при H_1: зависимость нелинейная. При истинной нулевой гипотезе F-отношение

$$F_{расч} = \frac{(\text{вариация ошибки спецификации})/(K-2)}{(\text{необъясненная вариация})/(N-K)} \quad (9.1.8)$$

подчиняется F-распределению Фишера—Снедекора со степенями свободы $f_1 = K-2$ и $f_2 = N-K$.

Обработав данные, представленные в табл. 49, с помощью одной из предыдущих программ регрессионного анализа, получим следующее уравнение линейной регрессии:

$$Y = 4{,}8533333 \cdot X - 30{,}106666. \quad (9.1.9)$$

Проведя суммирование в табл. (9.1.7) с использованием значений $\hat{Y}_i$, полученных по уравнению (9.1.9), получаем следующее равенство:

$$12375{,}28 = 9716{,}37 + 2588{,}91 + 70{,}0.$$

Подставляя эти числовые данные в (9.1.8), получим $F_{расч}$-отношение

$$F_{расч} = \frac{2588{,}91/(10-2)}{70{,}0/(50-10)} = 184{,}92214,$$

что больше, чем $F_{0,05;\,8:\,40}=2{,}18$. Это означает, что нулевая гипотеза с $P_{дов}=95\%$ отвергается и, следовательно, необходимо специфицировать зависимость $Y=f(X)$ как нелинейную. Отношение объясненной вариации к общей вариации, которое является коэффициентом детерминации, равно

$$r^2 = \frac{\sum_{i=1}^{k} n_i (\hat{Y}_i - \bar{Y}_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2}. \qquad (9.1.10)$$

Разделив обе части выражения (9.1.7) на общую вариацию и выполнив несложные преобразования, имеем

$$\hat{\eta}^2 = r^2 + \frac{\sum_{i=1}^{k} n_i (\bar{Y}_i - \hat{Y}_i)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{.,})^2}. \qquad (9.1.11)$$

Тогда выражение (9.1.4) для F-отношения можно записать в следующем виде:

$$F = \frac{(\hat{\eta}^2 - r^2)/(K-2)}{(1-\hat{\eta}^2)/(N-K)}. \qquad (9.1.12)$$

Обозначим отношение сумм квадратов в (9.1.11)

$$\frac{\sum_{i=1}^{k} n_i (\bar{Y}_i - \hat{Y}_i)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2} = B^2. \qquad (9.1.13)$$

Величина B^2 представляет собой отношение вариации ошибки спецификации к общей вариации и является показателем криволинейности зависимости

$$B^2 = \hat{\eta}^2 - r^2. \qquad (9.1.14)$$

Увеличение криволинейности приводит к уменьшению значения r^2 и соответственно к возрастанию показателя криволинейности B^2. С использованием программы П 47/34 производится полный анализ табличных данных, в результате которого определяются следующие величины: квадрат корреляционного отношения, коэффициент корреляции, показатель криволинейности, критерий Фишера, константы для прямого и обратного уравнений регрессии, стандартные отклонения для этих уравнений, доверительные интервалы и ошибки аппроксимации. Обрабатываемую таблицу данных в этом случае удобно представить следующим образом (табл. 50).

Таблица 50

Значения переменной X	Значения переменной Y				
X_1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1n_1}
X_2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2n_2}
X_3	Y_{31}	Y_{32}	Y_{33}	...	Y_{3n_3}
...	...	...	...	...	...
$X_{к-2}$	$Y_{к-2;\,1}$	$Y_{к-2;\,2}$	$Y_{к-2;\,3}$	...	$Y_{к-2;\,n_{к-2}}$
$X_{к-1}$	$Y_{к-1;\,1}$	$Y_{к-1;\,2}$	$Y_{к-1;\,3}$	...	$Y_{к-1;\,n_{к-1}}$
$X_к$	$Y_{к;\,1}$	$Y_{к;\,2}$	$Y_{к;\,3}$	...	$Y_{к;\,n_к}$

Число наблюдений за величиной Y при одном и том же значении X, n_1; n_2; ...; $n_к$, не обязательно должно быть одним и тем же, оно может меняться при изменении значения X, как это было и ранее в дисперсионном анализе.

Пример. В качестве исходных данных используем результаты, приведенные в табл. 49. Требуется провести полный корреляционно-регрессионный анализ этих данных.

Решение. Используя программу П 47/34, получаем следующие результаты анализа.

Результаты по П 47/34—1: $\hat{\eta}^2=9{,}9434355\cdot10^{-1}$; $f_2=40$; $F_{\hat{\eta}^2}=781{,}2876$.

По окончании работы П 47/34—1 имеем следующие промежуточные результаты во всех регистрах ПМК:

— операционный стек: ПХ=781,2876, ПУ=1367,2533, ПZ=1367,2533, ПТ=1367,2533;

— регистры прямого доступа:

П9=114166, П8=50, П7=114096; П6=—101790,72 П5=0; П4=1367,2533; П3=$9{,}9434355\cdot10^{-1}$; П2=70; П1=14; П0=40; ПА=2256; ПВ=12425; ПС=775; ПД=36970.

Результаты по П 42/34—2:

$r=8{,}860834\cdot10^{-1}$; $\bar{Y}_{..}=45{,}12$; $s_{Y/x}=7{,}3663677$;

$B^2=2{,}0919976\cdot10^{-1}$; $s_Y=15{,}892033$;

$F_{B_2}=4{,}8683668$ $a_{Y/x}=4{,}8533333$;

$\bar{X}_{.}=15{,}5$; $b_{Y/x}=-30{,}106666$;

$s_x=2{,}9014422$; $b_{x/Y}=8{,}200752$; $a_{x/Y}=1{,}6177411\cdot10^{-1}$.

9.2. Аппроксимация нелинейной зависимости экспонентой

Весьма большой класс задач обработки экспериментальных данных аппроксимируется экспоненциальным уравнением типа

$$Y=b_0\cdot e^{b_1\cdot x}\ .$$

В частности, такое уравнение используется в химической кинетике (уравнение Аррениуса), в идентификации объектов управления по динамическим переходным характеристикам и определении коэффициентов передаточных функций, например методом последовательного логарифмирования. К аналогичным уравнениям приводит решение многих задач исследования переходных процессов в электрических цепях.

Для аппроксимации экспериментальных данных экспонентой с помощью ПМК 34 можно использовать ранее приведенную программу линейного регрессионного анализа П 43/34 с некоторыми изменениями в инструк-

ции к программе. В библиотеке программ дается инструкция к П 43/34 для аппроксимации данных экспонентой [7].

9.2.1. Оценка точности аппроксимации зависимостей нелинейным уравнением

Основным показателем точности и качества аппроксимации экспериментальных данных той или иной зависимостью является величина среднего квадрата ошибки, вычисляемая по формуле

$$s^2 = \frac{1}{n-K-1}\sum_{i=1}^{n}[Y_i - \hat{Y}_i)^2 =$$

$$= \frac{1}{n-K-1}\sum_{i=1}^{n}[Y_i - f(X_i)]^2, \qquad (9.2.1)$$

где n — объем выборки (число наблюдений); K — число переменных в аппроксимирующем уравнении f(X); $\hat{Y}_i$ — оценка значения Y по уравнению f(X). Статистика s называется стандартной ошибкой оценки. Чем точнее оценки $\hat{Y}_i$ по найденной аппроксимирующей функции $Y=f(X)$ к фактически наблюдаемому значению Y_i, тем ближе к нулю будут располагаться оценки s. Это позволяет сравнивать точность аппроксимирующих функций путем сравнения величины s, получаемой для одной и той же выборки, но с использованием разных зависимостей. Выбирают ту зависимость, которая обеспечивает минимум статистики s.

Для уравнений типа $Y=b_0+b_1\cdot X$, $Y=b_0\cdot e^{b_1 x}$ и $Y=b_0\cdot X^{b_1}$ $K=1$, для уравнения $Y=ax^2+bx+c$ $K=2$ и т. д.

Нередко вместо абсолютного значения среднего квадрата ошибки вычисляют его относительное значение

$$s_{отн}^2 = \frac{1}{n-K-1}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{Y}_i}{Y_i}\right]^2. \qquad (9.2.2)$$

Программа П 49/34 вычисляет значение s или $s_{отн}$ для найденной аппроксимирующей функции f(X).

9.2.2. Повышение точности аппроксимации зависимости экспонентой

Поскольку при аппроксимации экспериментальных данных экспонентой происходит сглаживание не самой функции Y, а ее логарифма lnY, то первично полученные значения для b_0 и b_1 могут быть не оптимальны, т. е. не обеспечивают минимум стандартной ошибки оценки s. Для определения оптимальных значений b_0 и b_1 можно воспользоваться методом итерационных приближений, разработанным Ю. В. Линником. Оптимизация значений b_0 и b_1 производится с использованием рекуррентных соотношений [7]. В качестве начального нулевого приближения берутся значения b_0 и b_1, полученные с помощью логарифмирования по программе П43/34.

Для итерационной оптимизации значений b_0 и b_1 на ПМК 34 используется программа П 50/34 [7]. Итерации следует повторять до тех пор, пока получаемые значения среднего квадрата ошибки s^2_j не перестанут изменяться.

9.3. Аппроксимация зависимости степенным уравнением

Достаточно часто при аппроксимации экспериментальных данных эмпирической зависимостью используется зависимость $Y=b_0 \cdot X^{b_1}$, где b_0 и b_1 — нецелые числа. Логарифмируя это выражение, мы приходим к обычной линейной зависимости, что позволяет нам использовать программу П 43/34. Инструкция к программе П 43/34 при аппроксимации экспериментальных данных степенной зависимостью [7] приведена в библиотеке программ. Итерационную оптимизацию значений b_0 и b_1 можно произвести с помощью программы П 50/34, в которую при наборе программы по адресу 15 вместо команды К НОП (код команды 54) следует набрать команду Fln (код команды 18). Инструкция к П 50/34 при этом остается без изменений.

Наименее формализуемым этапом при аппроксимации экспериментальных данных той или иной зависимостью является этап спецификации, идентификации вида этой зависимости. Например, мы установили, что иско-

мая зависимость нелинейная. Но класс нелинейных зависимостей очень велик. Как быть? Во многих случаях выбор типа, вида формулы зависимости может быть сделан на основе теоретического анализа характера изучаемой зависимости. Когда для построения теоретической модели зависимости недостаточно информации, можно подобрать формулу, сравнивая графики ряда формул с экспериментальными кривыми. Графики многих формул приведены в специальных справочниках.

ЗАДАЧИ К ГЛ. 9

1 [16, с. 247]. Провести регрессионный анализ зависимости степени диссоциации α йодистого водорода от температуры t °C, если известны следующие опытные данные:

t : 280 300 320 340 360 380 400

α : 0,178 0,182 0,186 0,191 0,196 0,202 0,213

t : 420 440 460 480

α : 0,213 0,220 0,228 0,236.

Для построения регрессионной модели второго порядка использовать программу П 48/34.

2[39, с. 450]. Провести аппроксимацию зависимости показаний влагомера Y (в делениях шкалы) от толщины образца X (см) древесины бука уравнением $Y = a \cdot X^b$.

X : 1 3 5 7 9

Y : 56 28 20 16 14.

3[57, с. 709]. При изучении скорости химической реакции получены данные, приведенные в табл. 51 (τ — время от начала опыта, Y — количество вещества в реакционной смеси к моменту τ).

Таблица 51

τ	3	6	9	12	15	18	21	24
Y	57,6	41,9	31,0	27,7	16,6	12,2	8,9	6,5

Используя программу П 43/34, определить константы уравнения регрессии $Y = b_0 \cdot e^{b_1 \tau}$.

4[57, с. 712]. В таблице приведены данные о зависимости коэффициента теплоотдачи Y (в ккал/м²·ч·град) горизонтальной стенки к кипящей воде от разности температур X стенки и кипящей воды.

X : 6,10 7,50 8,88 11,1 12,2

Y : 3185 5390 6860 10045 12740

Требуется аппроксимировать зависимость уравнением $Y = b_0 \cdot X^{b_1}$, а также провести итерационную оптимизацию этого уравнения.

5[43, с. 305]. В табл. 52 даны результаты серии опытов, в которых определялось уменьшение темпа размножения одного вида бактерий под действием рентге-

Таблица 52

Доза X, 10^3 Р	Остаточный уровень размножения Y, %									
1	94	96	97	92	95	93	96	94	95	
2	87	91	86	88	88	90	89	89	95	87
3	83	85	82	84	81	81	85	83		
4	77	71	77	79	76	78	75	79	75	
5	71	68	70	69	69	68	72			
6	63	66	64	64	63	65	67	65	68	62
7	62	58	60	59	64	60	61	63		

новского облучения; результаты выражены в процентах к уровню размножения необлученных бактерий. Требуется провести полный корреляционно-регрессионный анализ этих данных.

6[47, с. 229]. Для разработки норм погрузки животных в автомашины потребовалось установить нормы требуемой площади пола на 1 голову в зависимости от среднего веса животных в загружаемой группе. При проведении опытных перевозок были получены данные о весе перевозимых свиней и их количестве в кузове,

имеющем определенную площадь пола. При обработке материалов получены средние площади, приходившиеся на 1 голову при разном весе свиней (от 40 до 220 кг), приведенные в табл. 53.

Таблица 53

X, кг	220	200	180	160	140	120	100	80	60	40
Y, м²	0,60	0,52	0,54	0,45	0,46	0,41	0,35	0,27	0,22	0,19

Используя программы регрессионного анализа, аппроксимировать исследуемую зависимость одним из уравнений. Используя полученное уравнение, построить зависимость числа свиней, которых можно погрузить в кузов автомобиля ЗИЛ-150 (площадь кузова 7,97 м²), от их среднего веса.

7 [56, с. 89]. Определить зависимость размеров структурных блоков горной породы от глубины их положения в недрах. Как известно, устойчивость бортов карьеров существенно зависит от трещиноватости пород, слагающих борта карьера. При изучении трещиноватости пород действующего карьера Норильского комбината на различных горизонтах были определены средние размеры структурных блоков, образованных системами трещин. Результаты замеров блоков, по отдельным глубинам карьера, приведены в табл. 54.

Таблица 54

Глубина карьера X, десятки метров	Размеры структурных блоков по замерам Y м³								Число замеров
0,3	0,28	0,27	0,22	0,25	0,29	0,26	0,2		7
1,8	0,55	0,37	0,48	0,67	0,30	0,44	0,86	0,42	8
3,3	0,76	0,54	0,65	0,57	0,68	0,65	0,95		7
4,8	0,85	0,69	0,76	0,70	0,65	0,85	0,73	0,75	8
6,3	0,95	0,76	0,83	1,02	0,71	1,09	0,86		7
7,8	1,48	1,02	1,05	1,50	1,40				5
9,3	1,15	0,86	1,10	1,45	1,19	1,25			6
10,8	1,30	1,05	0,96	1,42	1,05	1,20			6
12,3	0,95	1,30	1,10	1,25	1,20				5
13,8	1,10	1,15	1,35	1,25	1,36	1,20			6

Провести корреляционно-регрессионный анализ данных при значении $\alpha = 5\%$.

8. Определить зависимость удельного объемного сопротивления кварца ρ (ом·см) от абсолютной температуры T (°K) (ρ перпендикулярно главной оси кристалла). Уравнение оптимизировать (табл. 55).

Таблица 55

ρ, ом·см	$2{,}5 \cdot 10^{16}$	$2 \cdot 10^{15}$	$1{,}3 \cdot 10^{14}$	$2 \cdot 10^{13}$	$1{,}2 \cdot 10^{12}$	$1{,}0 \cdot 10^{11}$	$8 \cdot 10^{9}$
T, °K	335	336	405	444	502	572	669

9[51, с. 24]. Определить зависимость площади листьев примулы от длины листовой пластинки в форме $Y = b \cdot X^{b_1}$.

Получив искомую зависимость, оценить стандартную ошибку оценки, провести итерационную оптимизацию уравнения (табл. 56).

Таблица 56

Длина листовой пластинки X, см	3	5	6	8
Физическая площадь листа Y, см²	7	20	30	55

10[52, с. 140]. Определить зависимость веса семян (Y) с одного растения от высоты растения кориандра X в форме $Y = b_0 \cdot X^{b_1}$ и $Y = aX^2 + bX + c$. Сравнить стандартные ошибки оценки s по обоим уравнениям (табл. 57).

Таблица 57

Y, г	24,0	17,6	13,2	14,4	11,2	8,0
X, см	112	95	78	61	44	27

11[52, с. 148]. Определить зависимость площади листа огурца (Y) от длины листа (X) (табл. 58).

Таблица 58

Y, см²	67,65	110,60	153,40	224,10	297,50	369,15	460,40	514,50
X, см	8	10	12	14	16	18	20	22

Зависимость аппроксимировать уравнениями $Y = b_0 \cdot X^{b_1}$ и $Y = aX^2 + bX + c$. Сравнить точность аппроксимации обо-

ими уравнениями. Какое уравнение дает более точные результаты при $8 \leqslant X \leqslant 20$?

12. На 12 участках одного и того же поля производились опыты с целью изучения зависимости урожайности пшеницы (Y) от глубины вспашки (X).

Аппроксимировать зависимость уравнением $Y = aX^2 + bX + c$ (табл. 59).

Таблица 59

Y, ц/га	20,8	24,7	22,4	26	27,8	30,2	31,0	26,2	24,8	22,8	25,4	23,2
X, см	8	9	9	10	8	11	12	13	14	15	13	9

13. Используя числовые данные задач 3, 8, 9 и 11, составьте условия задач по собственной тематике исследований. Нанесите экспериментальные точки и графики аппроксимирующих зависимостей на миллиметровую бумагу.

РЕШЕНИЯ

1. После ввода программы П 48/34 производим очистку регистров: БП 90 С/П (0) и переходим к началу программы: В/О С/П (0). Далее начинаем ввод таблицы: 280↑0,178 С/П (0,178) В/О С/П (0,178) 300↑0,182 С/П (0,182) В/О С/П (0,182) и т. д. ... 480↑0,236 С/П (0,236) БП 27 С/П ($a = 4{,}050332 \cdot 10^{-7}$) ИП1 ($b = -1{,}8280256 \cdot 10^{-5}$) ИПО ($C = 1{,}5093047 \cdot 10^{-1}$). Таким образом, искомое уравнение имеет следующий вид:

$$\alpha \approx 4{,}05 \cdot 10^{-7} \cdot t^2 - 1{,}828 \cdot 10^{-5} \cdot t + 1{,}509 \cdot 10^{-1}.$$

2. Вводим программу П 43/34 и производим очистку регистров: В/О С/П (0). Начинаем ввод данных: 1 Fln(0) 56 Fln(4,0253517) С/П (0) 3 Fln(1,0986122) 28 Fln (3,3322045) С/П (1,0986122) 5 Fln 20 Fln С/П 7 Fln 16 Fln С/П 9 Fln 14 Fln С/П (2,1972245) БП 37 С/П ($4{,}0255754 = \ln b_0$) $F e^x$ ($56{,}012514 = b_0$) С/П ($-6{,}3681591 \cdot 10^{-1} = b_1$).

Искомое уравнение имеет вид

$$Y \approx 56{,}013 \cdot X^{-0{,}6368}.$$

Проведите самостоятельно итерационную оптимизацию параметров b_0 и b_1 данного уравнения.

3. Для аппроксимации экспериментальных данных экспонентой пользуемся соответствующей инструкцией к программе П 43/34. После набора П 43/34 производим очистку регистров: В/О С/П (0). Начинаем ввод данных по п. 3 : 3↑57,6 Fln (4,0535225) С/П (3) 6↑41,9 Fln (3,7352858) С/П (6) и т. д. ... 24↑6,5 Fln (1,8718021) С/П Ввод данных окончен. Выполняем пп. 4,5 и 6 : БП 37 С/П (4,3639896) Fe^x (78,569957 $= b_0$) С/П ($-1{,}0368579 \cdot 10^{-1} = b_1$).

Таким образом, искомое экспоненциальное уравнение можно записать следующим образом:

$$Y \approx 75{,}57 \cdot e^{-0{,}1037 \cdot \tau}.$$

4. Искомая зависимость имеет следующий вид:

$$Y = 106{,}6302 \cdot X^{1{,}9067115}.$$

5. Вводим в ПМК программу П 47/34—1, при этом по адресу 74 набираем число 6 (7—1=6), а по адресу 78 — число 7, т. к. в опытах использовано 7 разных уровней облучения. По окончании ввода программы и перехода в режим работы начинаем ввод исходных данных: 1 П2 В/О С/П (0). По п. 3 инструкции вводим первую строку наблюдений: 94 С/П (1) 96 С/П (2) и т. д. 95 С/П (9). По п. 4 обработаем первую строку наблюдений: БП 37 С/П (94,666666 $= \bar{Y}_{.1}$). Далее по п. 5 С/П (0). Аналогично вводим оставшиеся 6 строк наблюдений.

После ввода всей таблицы выполняем п. 7: С/П ($9{,}7451662 \cdot 10^{-1} = \hat{\eta}^2$). Далее С/П ($54 = f_2$) С/П (344,1724 = F-отношение). Вводим программу 2, при ее вводе по адресу 46 набираем число 5 (7—2=5). Перейдя в режим автоматических вычислений, вычислим значение коэффициента корреляции: В/О С/П ($-9{,}8474866 \cdot 10^{-1} = r$). Далее С/П ($4{,}7867 \cdot 10^{-3} = B^2 = \hat{\eta}^2 - r^2$) С/П ($1{,}7078357 = F_{B_2}$) С/П ($3{,}9344262 = \bar{X}$) С/П ($2{,}0319846 = s_x$) С/П ($77{,}278688 = \bar{Y}_{..}$) С/П ($12{,}068044 = s_Y$) С/П ($-5{,}8484647 = a_{Y/x}$) С/П ($100{,}28904 = b_{Y/x}$) С/П ($-1{,}6580931 \cdot 10^{-1} = a_{X/Y}$) С/П ($16{,}747952 = b_{x/Y}$) 1↑9,8474866 /—/ ВП 1 /—/ FX^2—$F\sqrt{\ }$ 12,068044 × ($2{,}0996349 = s_{Y/x}$). Пункты 23—27 инструкции к П 47/34 предлагается выполнить самостоятельно.

Проведем краткий анализ полученных результатов. Проверим нулевую гипотезу $H_0: \hat{\eta}^2=0$. Поскольку гипотеза односторонняя, то возьмем в таблице [17, 18] значение $F_{0,05;\,6;\,54}=2,3$. Расчетное F-отношение 344,1724 гораздо больше табличного, что приводит к отклонению H_0, т. е. можно утверждать, что рентгеновское облучение влияет на уровень размножения бактерий. Проверим значимость показателя криволинейности зависимости $B^2=\hat{\eta}^2-r^2$ по F-критерию. Для этого показателя число степеней свободы $f_1=K-2=7-2=5$, а $f_2=N-K=61-7=54$. По таблицам находим $F_{0,05;\,5;\,54}=2,4$. Это позволяет нам отвергнуть гипотезу $H_1: B^2=\hat{\eta}^2-r^2>0$, т. е. гипотезу о криволинейном характере зависимости, т. к. $F_{табл}=2,4>1,7078357=F_{B_2}$. Значит исследуемая зависимость имеет линейный характер. Прямое уравнение регрессии имеет вид

$$Y=-5,8484647\cdot X+100,28904,$$

обратное

$$X=-1,6580931\cdot 10^{-1}\cdot Y+16,747952.$$

Коэффициент корреляции $r=-9,8474866\cdot 10^{-1}$, а стандартное отклонение при аппроксимации данных прямых уравнением регрессии $s_{Y/X}=2,0996349$.

6. Первоначально попытаемся идентифицировать вид искомой зависимости, исходя из содержательных соображений поставленной задачи. Очевидно, что искомая зависимость не является линейной. Это следует хотя бы из того, что животные, имеющие больший вес, имеют и больший рост. Из этого следует, что увеличение площади не может быть пропорционально весу, т. е. крупные животные на единицу веса требуют меньшую площадь пола, чем мелкие животные. Далее очевидно, что уравнение не должно содержать свободного члена, т. к. при весе, равном нулю, и требуемая площадь тоже равна нулю, т. е. $Y(0)\equiv 0$. Поскольку экспонента не удовлетворяет последнему условию $[Y(0)=b_0\cdot e^0=b_0\cdot 1=b_0\neq 0]$, то наиболее подходящим для данной задачи остается уравнение $Y=b_0\cdot X^{b_1}$. Используя программы П 43/34 и П 50/34, определим значения b_0 и b_1, а затем проведем их итерационную оптимизацию. Вводим программу

П 43/34 и согласно инструкции к этой программе при аппроксимации $Y=b_0\cdot X^{b_1}$ вначале производим очистку регистров, а затем ввод таблицы: В/О С/П (0) 220 Fln (5,3936275) 0,60 Fln ($-5,108257\cdot10^{-1}$) С/П (5,3936275) 200 Fln 0,52 Fln С/П (5,2983173) 180 Fln 0,54 Fln С/П (5,1929568) 160 Fln 0,45 Fln С/П (5,0751738) 140 Fln 0,46 Fln С/П (4,9416424) 120 Fln 0,41 Fln С/П (4,7874917) 100 Fln 0,35 Fln С/П (4,6051702) 80 Fln 0,27 Fln С/П (4,3820266) 60 Fln 0,22 Fln С/П (4,0943445) 40 Fln 0,19 Fln С/П (3,6888795) БП 37 С/П ($-4,3178689=\ln b_0$) Fe^x ($1,3328257\cdot10^{-2}=b_0$) С/П ($7,036949\cdot10^{-1}=b_1$).

Таким образом, в первом приближении искомое уравнение имеет следующий вид:

$$Y\approx1,33\cdot10^{-2}\cdot X^{0,704}.$$

Вводим далее программу П 50/34 и начинаем уточнение искомой зависимости: 1,3328257 ВП 2 /—/ 7,036949 ВП 1 /—/ В/О С/П ($1,3328257\cdot10^{-2}$) С/П (0) 220↑0,6 С/П (0,6) 200↑0,52 С/П (0,52) 180↑0,54 С/П 160↑0,45 С/П 140↑0,46 С/П 120↑0,41 С/П 100↑0,35 С/П 80↑0,27 С/П 40↑0,19 С/П (0,19) БП 61 С/П ($4,7566326\cdot10^{-4}=s^2$) С/П ($1,374755\cdot10^{-2}=b_0$) ИПВ ($6,9772027\cdot10^{-1}=b_1$). Проведя несколько итерационных циклов, получим следующие результаты (табл. 60).

Таблица 60

Номер цикла j	s^2_j	b_0	b_1
0	$4,7566326\cdot10^{-4}$	$1,374755\cdot10^{-2}$	$6,9772027\cdot10^{-1}$
1	$4,7449127\cdot10^{-4}$	$1,374774\cdot10^{-2}$	$6,9781312\cdot10^{-1}$
2	$4,7444948\cdot10^{-4}$	$1,3747859\cdot10^{-2}$	$6,9781138\cdot10^{-1}$

Поскольку изменения происходят только в 4—5-м знаке после запятой, то дальнейшие циклы не имеет смысла проводить и можно за искомое уравнение взять следующее выражение:

$$Y\approx1,375\cdot10^{-2}\cdot X^{0,6978}.$$

Используя данное выражение, получаем формулу для нахождения числа животных, которые могут быть загружены в автомашину с площадью пола А: $n=A/Y$,

в частности для автомашины ЗИЛ-150, имеющей площадь 7,97 м², имеем формулу

$$n = \frac{7{,}97}{1{,}375 \cdot 10^{-2} \cdot X^{0{,}6978}}.$$

Составим программу нахождения n по X. Используемые константы занесем в регистры: 7,97 П 7 1,372 ВП 2 (—) П8 0,6978 П9. Программа в этом случае будет иметь вид ИП9⇄FX^Y ИП8 × ИП7 ÷ С/П БП 00. Последние две команды указывают на то, что программа начинается с адреса 00. Очевидно, что для нахождения n необходимо набрать значение среднего веса $\bar{X}$, затем выполнить В/О С/П (при первом вычислении n); результат n округлить до целых. При повторных вычислениях n используется одна команда С/П. Используя эту программу, вычислим n при $\bar{X}=90$: 90 В/О С/П ($25{,}143972 \approx 25 = n$), если $\bar{X}=140$, тогда 140 С/П ($18{,}472945 \approx 18 = n$) и т. д.

7. Используем программу П 47/34. По окончании работы П 47/34—1 имеем следующие результаты (без указания их расположения по регистрам памяти): 55; 14; 1,192056; $8{,}5798815 \cdot 10^{-1}$; $8{,}0022288 \cdot 10^{-1}$; 0; —49,111538; 56,313544; 65; 57,5056; 56,5; 3987; 426; 456,015; 36,921301; $8{,}0022288 \cdot 10^{-1}$; $8{,}5798815 \cdot 10^{-1}$; $8{,}5798815 \cdot 10^{-1}$.

Обратим внимание на то, что в данной задаче K=10, поэтому при наборе П 47/34—1 по адресу 78 вместо K=10 набираем команду С/П. При останове по этому адресу (п. 1 инструкции) выполняем команду ↑, затем набираем число 10 (K) и продолжаем вычисление. При этом на данном отрезке программы выполняется нахождение f=N—K, т. е. 65—10=55. Выполнение команды ↑ обязательно, в противном случае при наборе числа 10 происходит стирание числа 65 из регистра X без его пересылки в регистр Y.

После продолжения вычислений по П 47/34—2 по п. 22 получаем следующие числовые данные (идентифицируйте их как искомые параметры): $1{,}2545179 \cdot 10^{-1}$; $8{,}5588338 \cdot 10^{-1}$; 6,5538461; 3,224667; $8{,}6923076 \cdot 10^{-1}$; 4,3212077; $7{,}1730776 \cdot 10^{-2}$; $3{,}6215633 \cdot 10^{-1}$; 10,212302; $3{,}9911829 \cdot 10^{-1}$; —2,3230009.

Выполнив обработку исходной таблицы данных по П 47/34 и проверив правильность результатов по приве-

денным выше числам, сформулируйте выводы относительно результатов корреляционно-регрессионного анализа при $\alpha=5$. Можно ли ограничиться линейной регрессионной моделью или же требуется использование нелинейной модели?

8. С помощью программы П 43/34 находим аппроксимирующее уравнение в виде

$$\rho = 2542{,}3742 \cdot e^{\frac{10030{,}981}{T}}.$$

Далее, используя программу П 50/34, оптимизируем параметры этого уравнения, уменьшая при этом дисперсию оценки. Исходная дисперсия равна $6{,}4206474 \cdot 10^{28}$, конечная — $3{,}0742888 \cdot 10^{25}$. Как видим, это существенное улучшение аппроксимации. Определите значения b_0 и b_1, при которых достигается такое оптимальное значение дисперсии.

9. С помощью программы П 43/34 находим первое приближение искомого уравнения

$$Y = 6{,}8977918 \cdot 10^{-1} \cdot X^{2{,}1029305}.$$

Далее, используя П 50/34 (по адресу 15 вместо команды К НОП с кодом 54 набираем команду Fln с кодом 18), производим оптимизацию параметров этого уравнения. Вследствие этого дисперсия оценки снижается от $6{,}1546125 \cdot 10^{-2}$ до $2{,}8942947 \cdot 10^{-2}$. Каковы же оптимальные значения параметров b_0 и b_1? Ниже приведены значения b_0 и b_1 без указания номера итераций:

$6{,}6170176 \cdot 10^{-1}$;	2,1257984;
$6{,}6170738 \cdot 10^{-1}$;	2,1257941;
$6{,}6170168 \cdot 10^{-1}$;	2,1257984;
$6{,}6120073 \cdot 10^{-1}$;	2,1257422.

10. Применив программы П 43/34 и П 48/34, находим значения b_0, b_1, a, b и c для первого и второго уравнений. Идентифицируйте приведенные ниже числа как параметры этих уравнений: 8,4108486; $8{,}4045111 \cdot 10^{-1}$; $1{,}2852134 \cdot 10^{-2}$; $-1{,}393878 \cdot 10^{-2}$; $6{,}7814918 \cdot 10^{-1}$. Далее используем программу П 49/34, с помощью которой определяем два стандартных значения: 1,8880984 и

1,5506681. Какому уравнению отвечает меньшее значение полученного стандарта оценки, т. е. какое уравнение дает более точную аппроксимацию? Как определить, достоверно ли отличие этих двух оценок?

Обратите внимание на то, что при использовании программы П 49/34 для вычисления среднего квадрата ошибки аппроксимации по уравнению $Y=aX^2+bX+c$, т. е. при $K=2$, после ввода первой пары чисел наблюдается аварийный останов: ЕГГОГ, т. е.: В/О 1 ПО 112 С/П ($22{,}971422=\hat{Y}_1$) 24 С/П (ЕГГОГ). Внимательно подумайте над причиной этого, проанализировав часть программы П 49/34 с адреса 15 по адрес 17. Останов происходит при выполнении команды по адресу 17. Для определения содержимого регистров X и Y выполните после останова команды ⇄ ⇄. Как продолжить вычисления далее?

11. С помощью программы П 43/34 и П 48/34 находим значения b_0, b_1, a, b и c для первого и второго уравнения. Идентифицируйте приведенные ниже числа как параметры этих уравнений: —82,320957; $7{,}000103\cdot10^{-1}$; $9{,}781606\cdot10^{-1}$; 12,318728; 2,0476351. Далее с помощью программы П 49/34 находим значения стандартных отклонений оценки по этим уравнениям: $s=13{,}6218$ и $s=9{,}8408956$. Какому уравнению отвечает меньшее значение стандарта? Определите, какое уравнение дает более точные оценки при $8\leqslant X\leqslant 20$? Замечания относительно аварийного останова (ЕГГОГ) даны в задаче 10.

12. Идентифицируйте приведенные ниже числа как параметры уравнения $Y=aX^2+bX+c$: 9,878485; —28,868022; $-4{,}2967575\cdot10^{-1}$.

ГЛАВА 10

ДВУХМЕРНОЕ НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Из наблюдений установлять теорию, через теорию исправлять наблюдение — есть лучший способ к изысканию правды.

М. В. Ломоносов

Наряду с одномерным нормальным законом распределения в практике статистического анализа экспериментальных данных очень часто приходится иметь дело с нормальным законом распределения двух и большего числа измерений — так называемые нормальные многомерные распределения. Нередко в таком наборе переменных нельзя выделить однозначно независимые, входные переменные X_i и зависимые, выходные переменные Y_i. Рассмотрим основные особенности двухмерного нормального закона распределения вероятности и методы решения некоторых задач, связанных с такими распределениями.

Двухмерный нормальный закон распределения выражается формулой

$$f(X,Y) = \frac{1}{2\pi\sqrt{1-\rho^2}\,\sigma_X\sigma_Y}\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\times\right.$$

$$\times\left[\frac{(X-\bar{X}_0)^2}{\sigma^2_X}+\frac{(Y-\bar{Y}_0)^2}{\sigma_Y}-2\cdot\frac{X-\bar{X}_0}{\sigma_X}\times\right.$$

$$\left.\left.\times\frac{Y-\bar{Y}_0}{\sigma_Y}\cdot\rho\right]\right\}, \qquad (10.0.1)$$

где $f(X, Y)$ — плотность вероятности в точке X, Y. Как видно из (10.0.1), двухмерный нормальный закон распределения определяется пятью параметрами:

$\bar{X}_0$ — генеральное среднее переменной X;

σ^2_x — дисперсия переменной X;

$\bar{Y}_0$ — генеральное среднее переменной Y;

σ^2_Y — дисперсия переменной Y;

ρ — коэффициент корреляции переменных X и Y.

Преобразование переменных:

$$t_X = \frac{X - \bar{X}_0}{\sigma_X}; \quad t_Y = \frac{Y - \bar{Y}_0}{\sigma_Y}; \quad \varphi(t_X, t_Y) = f(X, Y) \cdot \sigma_X \sigma_Y$$

приводит уравнение нормальной корреляции (10.0.1) к стандартному виду

$$\varphi(t_X, t_Y) = \frac{(\sqrt{1-\rho^2})^{-1}}{2\pi} \exp\left[-\frac{1}{2(1-\rho^2)} \times \right.$$
$$\left. \times (t_X^2 + t_Y^2 - 2t_X t_Y \rho)\right]. \qquad (10.0.2)$$

Уравнению (10.0.2) в декартовой системе координат отвечает так называемая поверхность двухмерного нормального распределения (рис. 17).

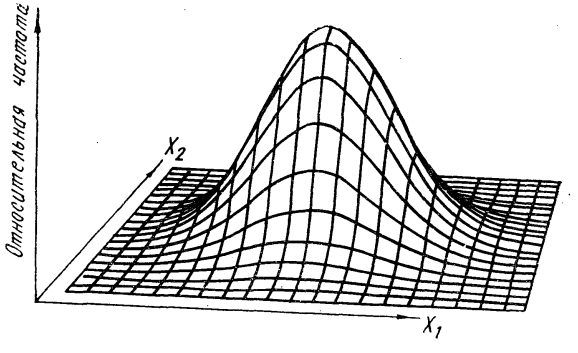

Рис. 17. Поверхность двухмерной нормальной плотности распределения вероятности

10.1. Корреляционный эллипс и двухмерная доверительная область

Анализ выражения (10.0.1) позволяет сделать следующие выводы:

1. Сечения нормальной поверхности распределения плоскостями, параллельными координатным плоскостям (X0φ) и (Y0φ), представляют собой обычные кривые плотности одномерного нормального распределения.

2. Условные математические ожидания $\bar{Y}_X$ и $\bar{X}_Y$ для этих законов рассчитываются по формулам:

$$\left.\begin{aligned} \bar{Y}_X &= \bar{Y}_0 + \rho\,\frac{\sigma_Y}{\sigma_X}\cdot(X - \bar{X}_0)\,, \\ X_Y &- \bar{X}_0 + \rho\,\frac{\sigma_X}{\sigma_Y}\cdot(Y - \bar{Y}_0)\,, \end{aligned}\right\} \qquad (10.1.1)$$

что означает, что линии регрессии Y по X и X по Y есть прямые линии.

3. Условные дисперсии $\sigma^2_{Y.X}$ и $\sigma^2_{X.Y}$ рассчитываются по формулам:

$$\left.\begin{aligned} \sigma^2_{Y.X} &= \sigma^2_Y\cdot(1 - \rho^2)\,, \\ \sigma^2_{X,Y} &= \sigma^2_X\cdot(1 - \rho^2)\,, \end{aligned}\right\} \qquad (10.1.2)$$

т. е. условные дисперсии постоянны.

4. Сечения нормальной поверхности распределения плоскостями, параллельными координатной плоскости X0Y, образуют в проекции на данную плоскость семейство концентрических эллипсов, определяемых уравнением

$$\frac{1}{1-\rho^2}\left\{\frac{(X-\bar{X}_0)^2}{\sigma^2_X} + \frac{(Y-\bar{Y}_0)^2}{\sigma^2_Y} - 2\,\frac{X-\bar{X}_0}{\sigma_X}\times\right.$$
$$\left.\times\,\frac{Y-\bar{Y}_0}{\sigma_Y}\cdot\rho\right\} = \chi^2, \qquad (10.1.3)$$

где χ^2 — параметр эллипса. Получаемые при этом эллипсы называются корреляционными. На рис. 18, *а* показано одно такое сечение, дающее корреляционный эллипс рис. 18, *в*. Центр координат в данном случае находится в точке $\bar{X}_0 = \mu_1$; $\bar{Y}_0 = \mu_2$.

На контуре корреляционного эллипса плотность вероятности постоянна и равна

$$f(X, Y) = \frac{1}{2\pi\,\sigma_X\,\sigma_Y\,\sqrt{1-\rho^2}}\cdot e^{-\chi^2/2}. \qquad (10.1.4)$$

В каждой из точек, находящихся внутри такого эллипса, плотность вероятности больше, и в каждой из точек, находящихся вне эллипса, плотность вероятности

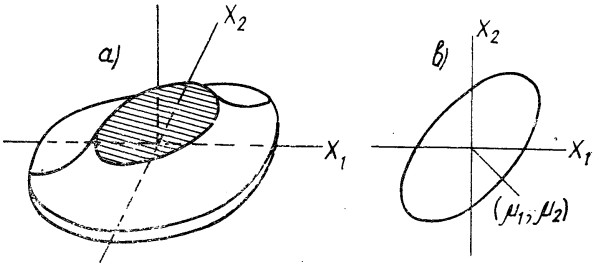

Рис. 18. Сечение поверхности двухмерной нормальной плотности вероятности плоскостью постоянной плотности (*а*) и соответствующий этому сечению корреляционный эллипс (*в*)

меньше, чем на контуре эллипса. Суммарная вероятность для всех точек, лежащих внутри эллипса, равна

$$P_{\text{дов}} = P_{\text{осн}} = 1 - e^{-\chi^2/2} \qquad (10.1.5)$$

и называется основной вероятностью корреляционного эллипса. Суммарная вероятность для всех точек, лежащих вне эллипса, равна $P_{\text{доп}} = e^{-\chi^2/2}$. Отсюда $\chi^2 = -2\ln[1 - P_{\text{дов}}]$. Основная вероятность $P_{\text{осн}}$ корреляционного эллипса больше аналогичной суммарной вероятности точек, лежащих внутри любой другой равновеликой эллипсу фигуры корреляционного поля. И наоборот, если задана величина $P_{\text{осн}}$, то наименьшая площадь ограничивается корреляционным эллипсом. Именно эти особенности делают очень удобным графическую характеристику двухмерных распределений с помощью такого эллипса.

В главе 3 мы рассмотрели методы построения доверительных и толерантных областей для одномерных величин. В случае же построения аналогичных доверительных областей для двухмерных величин — векторов, эти методы уже применять нельзя. Рассмотрим далее методы построения таких доверительных областей для двухмерных векторов, используя особенности корреляционных эллипсов.

Приведем несколько примеров, иллюстрирующих возможности излагаемой далее программы построения корреляционного эллипса.

1. Разработанный прибор характеризуется двумя закоррелированными показателями качества. Известны все пять параметров данного двухмерного распределения. Задав $P_{осн}=50\%$, требуется построить корреляционный эллипс, определяющий область расположения значений показателей качества для половины приборов.

2. Разработан процесс химической технологии. Определены параметры двухмерного нормального распределения. Задав $P_{осн}=95\%$, требуется построить корреляционный эллипс, а затем установить допуски для входной переменной X.

3. Имея все параметры двухмерного распределения закоррелированных переменных X и Y, определить долю выпускаемой продукции, удовлетворяющей одновременно двум ограничениям на эти показатели качества.

4. Имея корреляционный эллипс для двух показателей качества X и Y, произвести его аппроксимацию прямоугольной фигурой.

5. Зная параметры двухмерного нормального распределения, построить доверительную область для двухмерного вектора средних на основе имеющейся выборки наблюдений и заданной доверительной вероятности.

Для иллюстрации рассмотренных положений обратимся к примеру корреляционной зависимости между двумя механическими характеристиками стали 30 ХГСА — пределом прочности ($\sigma_в$) и пределом текучести (σ_s) [28]. Ниже в корреляционной таблице приведена так называемая двухмерная диаграмма распределения 209 плавок стали 30 ХГСА по $\sigma_в$ и σ_s. Значения $\sigma_в$ и σ_s в таблице являются серединами интервалов, в которые относятся наблюдения (табл. 61).

Таким образом, 209 наблюдений распределены по 72 клеткам весьма неравномерно. Например, в прямоугольник с серединой интервала $\sigma_s=92{,}75$ кг/мм^2 и серединой интервала $\sigma_в=102{,}25$ кг/мм^2 попало только 2 наблюдения, а для клетки с $\sigma_в=122{,}25$ кг/мм^2 и $\sigma_s=112{,}75$ кг/мм^2 таких наблюдений 31. То есть плотность наблюдений в центре эллипса выше, чем на его периферии. Основные параметры этого двухмерного распределения имеют следующие значения: $\bar{X}=119{,}4$; $\sigma_x=5{,}84$; $\bar{Y}=110{,}83$; $\sigma_Y=6{,}46$; $\rho=0{,}882$, где через X обозначен предел прочности $\sigma_в$, а через Y — предел текучести σ_s.

Таблица 61

σ_s, кг/мм²	σ_B, кг/мм²									
	102,25	106,25	110,25	114,25	118,25	122,25	126,25	130,25	134,25	Итого
127,75									1	1
122,75							2	6	2	10
117,75						19	20	2		41
112,75				1	26	31	8			66
107,75			4	21	26	3				54
102,75			8	15	1	1				25
97,75		1	5	2						8
92,75	2	2								4
Итого...	2	3	17	39	53	54	30	8	3	209

Зная эти параметры, можно оценить и теоретическое значение числа наблюдений для любой интервальной клетки с координатами центра клетки X_0 и Y_0. Возьмем, к примеру, интервальную клетку с $X_0=122{,}25$ и $Y_0=112{,}75$. Вычислим согласно (10.1.3) значение χ^2: $\chi^2=3{,}1806732\cdot10^{-1}\approx0{,}3181$. Затем, подставляя значение χ^2 в (10.0.2), вычислим значение $\varphi(t_x;\ t_Y)=2{,}8807255\cdot10^{-1}\approx0{,}2881$. Приближенно частоту попадания наблюдений в клетку можно вычислить по следующей формуле:

$$m_{теор}\approx\varphi(t_X;t_Y)\cdot\frac{i_X\cdot i_Y}{\sigma_X\cdot\sigma_V}\cdot n\,, \qquad (10.1.6)$$

где i_x и i_Y — длина интервала клетки по осям X и Y. Отсюда находим значения $m_{теор}$:

$$m_{теор}=\frac{0{,}2881\cdot4\cdot5\cdot209}{5{,}84\cdot6{,}46}=31{,}918\approx32\,.$$

Фактическая же частота наблюдений в этой клетке, как видно из приведенной ранее корреляционной таблицы, равна 31. Получаемая оценка тем точнее, чем у́же интервалы i_x и i_Y. Программа П 51/34 позволяет решить эту прогнозную, интерполяционную задачу.

Пример. Используем данные рассмотренного выше примера с корреляционной таблицей по стали 30 ХГСА

[28] и вычислим $m_{теор}$ для клетки с $X_0=122{,}25$ и $Y_0=112{,}75$, $i_x=4$; $i_Y=5$; $n=209$.

Решение. После набора программы П 51/34 вводим исходные данные: 119,4 П0 5,84 П1 110,83 П2 6,46 П3 0,882 П4 122,25 П5 112,75 П6 4 П7 5 П8 209 П9. Далее В/О С/П ($31{,}917787=m_{теор}\approx 32$).

Очень часто кроме значений $m_{теор}$ для отдельных клеток корреляционного поля требуется построить и сам корреляционный эллипс, точки которого бы ограничивали доверительную область, содержащую определенную долю наблюдений.

10.2. Построение корреляционного эллипса

Для построения требуемых доверительных областей необходимо знание пяти основных параметров двухмерного нормального распределения: $\bar{X}$, s_x, $\bar{Y}$, s_Y и r. Далее, поскольку в уравнение корреляционного эллипса входит величина χ^2, необходимо задать и ее значение. Это значение вычисляется по формуле

$$\chi^2=-2\ln(1-P_{дов}), \qquad (10.2.1)$$

где $P_{дов}$ — значение доверительной вероятности, задаваемое самим исследователем.

Программа П 52/34 позволяет построить корреляционный эллипс в обычных, нестандартизованных переменных. Перед работой с программой П 52/34 необходимо построить оси X—Y и определить величину Y_{min}, равную значению переменной Y в точке пересечения оси. Кроме того, необходимо выбрать масштаб по оси Y. При построении эллипса на миллиметровой бумаге целесообразнее взять масштаб, относящийся к 1 мм. С 53-го по 59-й адрес при наборе программы П 52/34 набирается значение Y_{min}. Например, при $Y_{min}=3{,}69\cdot 10^{14}$ этот фрагмент программы приведен в табл. 62.

Таблица 62

Адрес	53	54	55	56	57	58	59
Команда	3	,	6	9	ВП	1	4
Код	03	0—	06	09	ОС	01	04

Если же показатель степени у числа Y_{min} отрицательный, например $Y_{min}=1{,}2\cdot10^{-25}$, тогда данный фрагмент программы будет выглядеть иначе (табл. 63).

Таблица 63

Адрес	53	54	55	56	57	58	59
Команда	1	,	2	ВП	2	5	/—/
Код	01	0—	02	ОС	02	05	0L

В том случае, когда после ввода очередного значения X_i и выполнения команд В/О С/П происходит останов с индикацией ЕГГОГ, то это означает, что прямая, проходящая через точку X_i на оси X и параллельная оси Y, не пересекается с доверительным эллипсом. Иными словами, точки с координатой по оси X, равной X_i, находятся левее или правее корреляционного эллипса.

Пример. При обработке данных по генераторным диодам сантиметрового диапазона были получены следующие результаты по частоте генерации и рабочему напряжению диодов: $\bar{f}=14{,}0$ ГГц; $\sigma_f=1{,}89$ ГГц; $\bar{u}_{раб}=7{,}0$ В; $\sigma_u=0{,}73$ В; $\rho=-0{,}72$.

Необходимо построить совместную доверительную область — корреляционный эллипс, для $P_{дов}=90\%$.

Решение. Определяем $\chi^2=-2\ln(1-0{,}9)=4{,}6051702\approx4{,}61$. Выбираем масштаб для оси f 0,08 ГГц в 1 мм. Набираем это число и заносим в регистр памяти, нажимая клавиши П Д. Строим на миллиметровой бумаге оси: f — вертикальная ось, $U_{раб}$ — горизонтальная ось. Точка пересечения осей (10,0; 5,5). В программу П 52/34 с 53-го по 56-й адрес заносим число 10,0, для чего нажимаем клавиши 10,0. Согласно инструкции к П 52/34 вводим в регистры памяти 5 параметров двухмерного нормального распределения: 2 средних, 2 стандарта и коэффициент корреляции, а также $\chi^2=4{,}61$.

Начинаем построение эллипса, для этого набираем число 5,5 — значение $U_{раб}$ в точке пересечения осей и нажимаем клавиши В/О С/П. После останова индицируется число 17,612972 (ГГц) $=Y_{1верх}$, координата пересечения эллипса с осью f. Нажимаем С/П — после останова индицируется число $l_{1верх}=95{,}16215$ (расстояние

от оси $U_{раб}$ до этой точки, мм). Нажимаем С/П, после останова индицируется число 15,979356 (ГГц) — координата нижней точки пересечения эллипса с осью f. Нажимаем С/П, после останова индицируется число 74,74195 — расстояние от оси $U_{раб}$ до этой точки. На горизонтальной оси $U_{раб}$ масштаб выбран 0,3 В в 1 см. Поэтому набираем число 5,8, нажимаем В/О С/П и определяем координаты следующих двух точек и т. д. Результаты расчетов приведены в табл. 64.

Соединяя точки с вычисленными координатами, получаем корреляционный эллипс, внутри которого находится 90% точек совокупности. На рис. 19 показаны эллипс и прямоугольник, построенный на основе доверительных интервалов для каждой переменной в отдельности с $P_{дов}=90\%$.

Сравнение этих двух фигур показывает, сколь велико отличие совместной двухмерной доверительной области — корреляционного эллипса от прямоугольника, полученного с помощью двух доверительных интервалов для каждой из переменных. Во-первых, верхняя и нижняя часть эллипса выходят из границы прямоугольника. Во-вторых, если для каждого из доверительных интервалов $P_{дов}=90\%$, то тогда, если переменные f и $U_{раб}$ независимые, согласно теории вероятностей доверительная вероятность прямоугольной совместной области равна $0{,}9^2=0{,}81$. В нашем же случае переменные f и $U_{раб}$ закоррелированы, т. е. не являются независимыми. В этом случае доверительная вероятность такой области (прямоугольника) $P_{дов}\geqslant 1-(1-0{,}9)-(1-0{,}9)=0{,}8$, более же точно определить $P_{дов}$ невозможно.

Проблема построения совместной доверительной области встречается и при совместном рассмотрении интервалов изменения коэффициентов регрессии в уравнении регрессии. В случае совместного рассмотрения изменения этих коэффициентов строится доверительный эллипс [29, с. 337].

Далее, при рассмотрении метода сравнения двух векторов средних и дискриминантного анализа, мы вновь встретимся с аналогичной проблемой построения доверительных областей (эллипсов).

Таблица 64

Параметры	Координаты точек, находящихся на корреляционном эллипсе					
$U_{раб}$, В	5,433	5,5	6,1	6,4	6,7	7,0
$f_{верх}$, ГГц	16,98278	17,612972	17,983306	17,720105	17,323315	16,816148
$l_{верх}$, мм	87,28475	95,16215	99,791325	96,501312	91,541437	85,201185
$f_{нижн}$, ГГц	16,895339	15,979356	13,372091	12,516826	11,795151	11,183852
$l_{нижн}$, мм	85,741737	74,74195	42,151137	31,460325	22,439387	14,79815
$U_{раб}$, В	7,3	7,6	7,9	8,2	8,5	8,567
$f_{верх}$, ГГц	16,204849	15,483174	14,627909	13,574709	12,020644	11,140661
$l_{верх}$, мм	77,560612	68,539675	57,848862	44,683862	25,25805	14,258262
$f_{нижн}$, ГГц	10,676685	10,279895	10,016694	9,951428	10,387028	11,01722
$l_{нижн}$, мм	8,4585625	3,4986875	0,208675	—0,60715	4,83785	12,71525

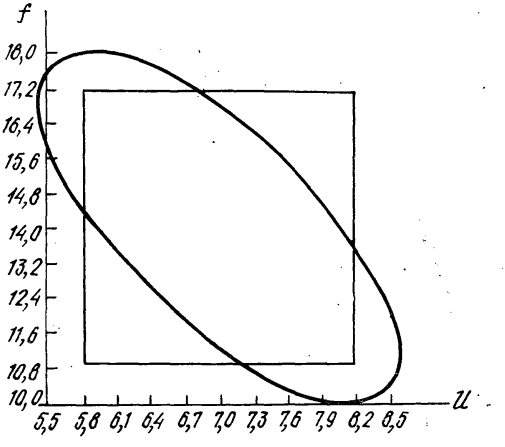

Рис. 19. Совместная доверительная область — корреляционный эллипс, — для вектора $[\bar{f}, \bar{U}]$ с $P_{дов}=90\%$

ЗАДАЧИ К ГЛ. 10

1[59, с. 83]. Исследовалась определенная категория новорожденных детей. При этом проводились измерения веса и роста при рождении. Средний вес новорожденных $\bar{X}=3509$ г, стандартное отклонение $s_x=462$ г. Средний рост $\bar{Y}=51{,}5$ см, стандартное отклонение $s_Y=2{,}21$ см. Коэффициент корреляции $r=0{,}804$. Требуется определить координаты точек корреляционного эллипса с шагом по оси X в 100 граммов и доверительной вероятностью $P_{дов}=50\%$. Использовать программу П 52/34.

2. Завод выпускает в течение нескольких лет прибор, характеризующийся двумя коррелированными параметрами: X и Y. Известно, что $\bar{X}=8{,}7$ условных единиц (у. е.), $\sigma_x=0{,}8$ у. е., $\bar{Y}=8{,}4$ у. е., $\sigma_Y=0{,}3$ у. е., $r_{xY}=-0{,}38$. В связи с расширением новой, нетрадиционной области использования этого прибора определилась новая группа потребления с параметрами: 8,6 у. е. $\leqslant X \leqslant$ 9,0 у. е. и 8,3 у. е. $\leqslant Y \leqslant$ 8,7 у. е. Требуется оценить процент приборов, удовлетворяющих этим ограничениям.

РЕШЕНИЯ

1. Примем за горизонтальную ось X вес новорожденных, а в качестве вертикальной оси Y рост новорожденного. Значение Y в точке пересечения с осью X—Y_{min} возьмем равным 50 см. Масштаб $m_Y = 1$ мм/мм. Согласно формуле (10.2.1) вычислим $\chi^2 = 1{,}3862944$. Введя все исходные данные в регистры памяти согласно п. 2 инструкции, а также набрав с 53-го по 57-й адрес значение $Y_{min} = 50{,}0$, начнем искать крайнее левое (минимальное) значение X, при котором имеются уже две координаты эллипса. Начинаем с X=2900 : 2900 В/О С/П (ЕГГОГ), далее берем 2950 В/О С/П (ЕГГОГ) 3000 В/О С/П (50,08817). Теперь возьмем X=2975 : 2975 В/О С/П (49,741034). Ограничимся этой точностью и начнем дальнейшие расчеты координат эллипса с 3000 г с шагом в 100 г. В табл. 65 даны округленные значения полученных результатов.

Таблица 65

X	2975	3000	3100	3200	3300	3400	3500
$Y_{верх}$	49,74	50,1	50,95	51,6	52,1	52,6	53,0
$l_{верх}$	—3	1,0	9,0	16,0	21,0	26,0	30,0
$Y_{низ}$	49,15	49,0	49,0	49,0	49,3	49,6	49,9
$l_{низ}$	—8	—10,0	—11,0	—10,0	—7,0	—4,0	—1,0

X	3600	3700	3800	3900	4000	4050	4052
$Y_{верх}$	53,4	53,7	53,9	54,1	54,1	53,7	53,68
$l_{верх}$	33,8	36,8	39,3	40,8	40,5	37,4	36,8
$Y_{низ}$	50,3	50,8	51,3	51,9	52,7	53,4	53,5
$l_{низ}$	3,2	7,9	13,1	19,3	27,2	34,2	34,96

Таким образом, половина новорожденных изучаемой категории будет иметь рост и вес, которые будут ограничены данным эллипсом. Постройте по найденным координатам корреляционный эллипс.

2. Применяем программу П 51/34. Поскольку $\Delta X = 0{,}4$ у. е. и $\Delta Y = 0{,}4$ у. е., то принимаем $i_x = i_Y = 0{,}4$. При этом $X_0 = 8{,}8$ у. е., $Y_0 = 8{,}5$ у. е. После набора П 51/34 вводим исходные данные: 8,7 П0 0,8 П1 8,4 П2 0,3 П3 0,38 /—/ П4 8,8 П5 8,5 П6 0,4 П7 П8 100 П9 В/О С/П (10,456626).

Таким образом, указанным ограничениям будет удовлетворять примерно 10,5% выпускаемых приборов.

ГЛАВА 11

МНОЖЕСТВЕННЫЙ ЛИНЕЙНЫЙ КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Комплексное изучение явления
ведет к открытию новых законов.

И. Л. Кнунянц

Линейную модель регрессионной зависимости с m независимыми переменными можно представить в виде

$$Y_i = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m + \varepsilon_i. \qquad (11.0.1)$$

Последовательность оценивания параметров уравнения множественной регрессии в принципе не отличается от последовательности при оценивании уравнения с одной переменной. Ниже мы рассмотрим модель регрессии для двух независимых переменных X_1 и X_2, программу, реализующую оценку параметров уравнения и другие процедуры, связанные с оценкой этого уравнения.

11.1. Основы анализа в матричном виде

Поскольку для нашего случая с двумя независимыми переменными система нормальных уравнений будет довольно громоздка, то рассмотрим модель линейной регрессии в матричном виде. Это будет полезно и по той причине, что с применением этого аппарата далее будут изложены методы сравнения векторов и метод дискриминантного анализа.

Изложение начнем с разбора примера, иллюстрирующего возможности множественного линейного регрессионного анализа [23, с. 63—79].

Исследовалось влияние коэффициента механизации работ и среднего возраста работников 14 предприятий на производительность труда. В качестве независимых переменных в данном случае выступают: коэффициент механизации работ X_1, %; средний возраст работников

X_2 лет. В качестве зависимой переменной — производительность труда в условных единицах измерения Y. В табл. 66 представлены полученные в результате исследования исходные данные.

Таблица 66

№ п/п	Расчетная производительность $\hat{Y}_i$, т/ч	Разность расчетной и фактической $Y_i - \hat{Y}_i = e_i$, т/ч	Фактическая производительность Y_i, т/ч	Коэффициент механизации работ X_1, %	Средний возраст работников X_2, лет
1	23,819256	—3,819256	20,0	32	33
2	22,449007	1,550993	24	30	31
3	27,196465	$8{,}03535 \cdot 10^{-1}$	28	36	41
4	28,981894	1,018106	30	40	39
5	30,622086	$3{,}77914 \cdot 10^{-1}$	31	41	46
6	33,30023	$-3{,}0023 \cdot 10^{-1}$	33	47	43
7	36,601145	—2,601145	34	56	34
8	36,185964	$8{,}14036 \cdot 10^{-1}$	37	54	38
9	39,978354	—1,978354	38	60	42
10	36,234377	3,765623	40	55	35
11	40,026766	$9{,}73234 \cdot 10^{-1}$	41	61	39
12	43,978334	$-9{,}78334 \cdot 10^{-1}$	43	67	44
13	44,393515	$6{,}06485 \cdot 10^{-1}$	45	69	40
14	48,234317	$-2{,}34317 \cdot 10^{-1}$	48	76	41

Определим следующие матрицы:

$$\mathbf{Y}^T = [Y_1\ Y_2\ \dots\ Y_{13}\ Y_{14}] \tag{11.1.1}$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1\,1} & X_{2\,1} \\ 1 & X_{1\,2} & X_{2\,2} \\ \vdots & \vdots & \vdots \\ 1 & X_{1\,13} & X_{2\,13} \\ 1 & X_{1\,14} & X_{2\,14} \end{bmatrix} \tag{11.1.2}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ X_{1\,1} & X_{1\,2} & \dots & X_{1\,13} & X_{1\,14} \\ X_{2\,1} & X_{2\,2} & \dots & X_{2\,13} & X_{2\,14} \end{bmatrix} \cdot \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{13} \\ Y_{14} \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^{14} Y_i \\ \sum\limits_{i=1}^{14} X_{1i}\ Y_i \\ \sum\limits_{i=1}^{14} X_{2i}\ Y_i \end{bmatrix} \tag{11.1.3}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ X_{1\,1} & X_{1\,2} & \dots & X_{1\,13} & X_{1\,14} \\ X_{2\,1} & X_{2\,2} & \dots & X_{2\,13} & X_{2\,14} \end{bmatrix} \cdot \begin{bmatrix} 1 & X_{1\,1} & X_{2\,1} \\ 1 & X_{1\,2} & X_{2\,2} \\ \vdots & \vdots & \vdots \\ 1 & X_{1\,13} & X_{2\,13} \\ 1 & X_{1\,14} & X_{2\,14} \end{bmatrix} \tag{11.1.4}$$

Знак Т означает транспонирование исходной матрицы. При заданных значениях X_{1i} и X_{2i} среднее $\hat{Y}_i$ зависит от параметров уравнения регрессии β_0, β_1 и β_2:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2, \tag{11.1.5}$$

где знак «^» показывает, что это оценка данного генерального параметра.

Запишем вектор искомых параметров регрессии следующим образом:

$$\widehat{\beta}^T = [\widehat{\beta}_0\ \widehat{\beta}_1\ \widehat{\beta}_2]. \tag{11.1.6}$$

Нормальные уравнения в матричной форме имеют вид

$$(\mathbf{X}^T\mathbf{X}) \cdot \widehat{\beta} = \mathbf{X}^T\mathbf{Y}. \tag{11.1.7}$$

Матрицу $(\mathbf{X}^T\mathbf{X})$, содержащую в себе исходную информацию о независимых переменных, называют информационной матрицей Фишера [27].

Решением уравнения (11.1.7) будет следующее выражение:

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1} \cdot \mathbf{X}^T\mathbf{Y} \tag{11.1.8}$$

(далее вместо $\widehat{\beta}$ пишем b). Матрицу, обратную информационной матрице Фишера, называют ковариационной матрицей.

Информационная матрица Фишера, симметричная и для нашего случая двух независимых переменных, имеет вид

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{2i} \\ \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X^2_{1i} & \sum_{i=1}^{n} X_{1i} X_{2i} \\ \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{1i} X_{2i} & \sum_{i=1}^{n} X^2_{2i} \end{bmatrix} \tag{11.1.9}$$

Из данных, имеющихся в табл. 66, получим следующие матрицы:

$$X^TX = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 32 & 30 & \dots & 76 \\ 33 & 31 & \dots & 41 \end{bmatrix} \cdot \begin{bmatrix} 1 & 32 & 33 \\ 1 & 30 & 31 \\ \vdots & \vdots & \vdots \\ 1 & 69 & 40 \\ 1 & 76 & 41 \end{bmatrix} =$$

$$= \begin{bmatrix} 14 & 724 & 546 \\ 724 & 40134 & 28533 \\ 546 & 28533 & 21544 \end{bmatrix}; \qquad (11.1.10)$$

$$X^TY = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 32 & 30 & \dots & 76 \\ 33 & 31 & \dots & 41 \end{bmatrix} \cdot \begin{bmatrix} 20 \\ 24 \\ \vdots \\ 45 \\ 48 \end{bmatrix} = \begin{bmatrix} 492 \\ 26907 \\ 19384 \end{bmatrix}. \qquad (11.1.11)$$

Для оценки уравнения регрессии (11.1.5) нам необходима матрица, обратная информационной $\mathbf{X}^T\mathbf{X}$, т. е. $(\mathbf{X}^T\mathbf{X})^{-1}$, которая для нашего примера имеет следующий вид:

$$(X^TX)^{-1} =$$

$$= \begin{bmatrix} 6{,}1678084 & -2{,}3001\cdot 10^{-3} & -1{,}5326748\cdot 10^{-1} \\ -2{,}300099\cdot 10^{-3} & 4{,}273464\cdot 10^{-4} & -5{,}0768753\cdot 10^{-4} \\ -1{,}5326748\cdot 10^{-1} & -5{,}0768753\cdot 10^{-4} & 4{,}6031328\cdot 10^{-3} \end{bmatrix}. \qquad (11.1.12)$$

Некоторые отличия в двух симметричных членах этой матрицы в четвертом знаке после запятой связаны с погрешностью обращения матрицы на микрокалькуляторе.

Далее согласно (11.1.8) получаем оценки $b = \hat{\beta}$:

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 6{,}1678084 & -2{,}3001\cdot 10^{-3} & -1{,}5326748\cdot 10^{-1} \\ -2{,}300099\cdot 10^{-3} & 4{,}273464\cdot 10^{-4} & -5{,}0768753\cdot 10^{-4} \\ -1{,}5326748\cdot 10^{-1} & -5{,}0768753\cdot 10^{-4} & 4{,}6031328\cdot 10^{-3} \end{bmatrix} \times$$

$$\times \begin{bmatrix} 492 \\ 26907 \\ 19384 \end{bmatrix} = \begin{bmatrix} 1{,}7361 \\ 5{,}25946\cdot 10^{-1} \\ 1{,}59178\cdot 10^{-1} \end{bmatrix}. \qquad (11.1.13)$$

Таким образом, выборочное уравнение регрессии в исходных переменных имеет следующий вид:

$$Y = 1{,}7361 + 5{,}25946 \cdot 10^{-1} \cdot X_1 + 1{,}59178 \cdot 10^{-1} \cdot X_2 \quad (11.1.14)$$

11.2. Точность и значимость оценок в уравнении регрессии

Помимо коэффициентов уравнения регрессии нас интересует также и точность оценки, обеспечиваемая этим уравнением. Согласно МНК вектор остатков равен

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb}. \quad (11.2.1)$$

Производя дальнейшее преобразование, получаем

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{N - m - 1} = \frac{\sum_{i=1}^{N} e_i^2}{N - m - 1}, \quad (11.2.2)$$

где $\mathbf{e}$ — вектор остатков, $e^2_i = (Y_f - \hat{Y}_i)^2$; N — общее число наблюдений (в нашем примере $N = 14$); m — число независимых переменных (в нашем примере $m = 2$). Выборочная статистика $\hat{\sigma}$ называется стандартной ошибкой оценки. Для нахождения $\hat{\sigma}$ определим по (11.1.14) значения $\hat{Y}_i$ и e_i, запишем их значения в первую и вторую колонки табл. 66. Далее согласно (11.2.2) находим $\hat{\sigma}^2$ и $\hat{\sigma}$. Согласно данным второго столбца табл. 66 имеем

$$\mathbf{e}^T\mathbf{e} = [-3{,}819256 \ldots -2{,}34317 \cdot 10^{-1}] \cdot \begin{bmatrix} -3{,}819256 \\ 1{,}550993 \\ 8{,}03535 \cdot 10^{-1} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ 6{,}06485 \cdot 10^{-1} \\ -2{,}34317 \cdot 10^{-1} \end{bmatrix} =$$

$$= (-3{,}819256)^2 + (1{,}550993)^2 + \ldots + (6{,}06485 \cdot 10^{-1})^2 + (-2{,}34317 \cdot 10^{-1})^2 = 46{,}756923. \quad (11.2.3)$$

Откуда $\hat{\sigma}^2 = 46{,}756923/(14-2-1) = 4{,}2506293$,

а $\hat{\sigma} = 2{,}0617054$.

Поскольку полученные коэффициенты регрессионного уравнения являются выборочными, то нас интересует точность этих оценок, а также их достоверность и значимость. Для решения этих задач используется ковариационная матрица векторной оценки **b**, которая имеет следующий вид:

$S_{\hat{\beta}} = \hat{\sigma}^2 \cdot (X^T X)^{-1}$, или в развернутом виде

$$S_b = \begin{bmatrix} S_{b_{00}} & S_{b_{01}} & S_{b_{22}} \\ S_{b_{10}} & S_{b_{11}} & S_{b_{12}} \\ S_{b_{20}} & S_{b_{21}} & S_{b_{22}} \end{bmatrix}. \qquad (11.2.4)$$

Корни квадратные из элементов главной диагонали матрицы S_b и являются стандартными ошибками коэффициентов регрессии b_0, b_1 и b_2. Для нашего примера имеем

$$S_b = 4{,}2506293 \times$$

$$\times \begin{bmatrix} 6{,}1678084 \\ -2{,}300099 \cdot 10^{-3} \\ -1{,}5326748 \cdot 10^{-1} \end{bmatrix} \begin{bmatrix} -2{,}3001 \cdot 10^{-3} \\ 4{,}273464 \cdot 10^{-4} \\ -5{,}0768753 \cdot 10^{-4} \end{bmatrix} \begin{bmatrix} -1{,}5326747 \cdot 10^{-1} \\ -5{,}0768753 \cdot 10^{-4} \\ 4{,}6031328 \cdot 10^{-3} \end{bmatrix}. \qquad (11.2.5)$$

Для проверки гипотез относительно величины коэффициента регрессионного уравнения пользуются выборочной статистикой типа

$$t = \frac{b_k - b_{0k}}{\sqrt{s_{b_{kk}}}}, \qquad (11.2.6)$$

где b_{0k} — предполагаемое значение коэффициента при X_k; $\sqrt{S_{b_{kk}}}$ — стандартная ошибка оценки коэффициента b_k. Нулевая гипотеза в этом случае формулируется так:

$$H_0 : b_k = b_{0k}.$$

Статистика (11.2.6) подчиняется t-распределению Стьюдента с числом степеней свободы $f = N - m - 1$.

Наиболее часто производится проверка нулевой гипотезы о равенстве коэффициента b_k нулю.

11.3. Мультиколлинеарность входных переменных

Для правильного и разумного использования такого способа проверки достоверности коэффициентов регрессии рассмотрим такое неприятное явление для регрес-

сионного анализа, как мультиколлинеарность независимых переменных [23]. Мультиколлинеарностью называется наличие линейной связи между независимыми переменными, в нашем случае X_1 и X_2.

Мультиколлинеарность может проявляться в двух видах: в функциональной зависимости между X_1 и X_2, например $X_2 = a_0 + b_1X_1$, и в стохастической, вероятностной. В том случае, когда $X_2 = a_0 + b_1X_1$, либо, наоборот, $X_1 = c_0 + b_2X_2$, определитель информационной матрицы $\mathbf{X}^T\mathbf{X}$ равен нулю, т. е. эта матрица вырожденная. Этот случай является нарушением предпосылок МНК в регрессионном анализе [23, 24]. Такой вид мультиколлинеарности, как правило, встречается крайне редко. Гораздо чаще мультиколлинеарность наблюдается в стохастической форме. Если в первом случае наблюдается точное равенство $|\mathbf{X}^T\mathbf{X}| = 0$ и, следовательно, информационная матрица не может быть обращена для применения выражения (11.1.8), то во втором случае наблюдается приблизительное равенство $|\mathbf{X}^T\mathbf{X}| \approx 0$. К чему приводит стохастическая форма мультиколлинеарности? Чем сильнее корреляция между X_1 и X_2, тем меньше определитель информационной матрицы, тем меньше точность оценки коэффициентов уравнения регрессии, тем сильнее взаимосвязь между коэффициентами b_0, b_1 и b_2. Эти оценки становятся ненадежными, а само уравнение регрессии теряет устойчивость. Неустойчивость уравнения выражается в том, что изменение объема выборки на 1, 2 или несколько наблюдений приводит к значительным сдвигам в оценках коэффициентов, которые могут увеличиваться, уменьшаться и даже менять знак.

Оценку силы мультиколлинеарности можно произвести, вычислив коэффициенты корреляции между коэффициентами b_i и b_j по формуле

$$r_{b_i b_j} = \frac{s_{b_{ij}}}{\sqrt{s_{b_{ii}} \cdot s_{b_{jj}}}}. \qquad (11.3.1)$$

В нашем примере имеем

$$r_{(b_0 b_1)} = -4{,}48 \cdot 10^{-2}; \quad r_{(b_0 b_2)} = -9{,}096 \cdot 10^{-1};$$

$$r_{(b_1 b_2)} = -3{,}62 \cdot 10^{-1}.$$

Очевидно, что для того, чтобы оценки коэффициентов b_i были независимыми, необходимо, чтобы в матри-

це Фишера (X^TX) только диагональные элементы были отличны от нуля. Для достижения этого используются так называемые ортогональные методы планирования экспериментов [27]. Какова цель получения взаимно независимых оценок коэффициентов? Во-первых, такие оценки имеют большую точность, т. е. значение S_b уменьшается. Во-вторых, если окажется, что одна или несколько оценок коэффициентов b_i недостоверны, то нет необходимости заново производить расчет остальных коэффициентов b_j.

В том случае, когда коэффициенты корреляции $r(b_ib_j)$ достаточно велики, выражение (11.2.6) не дает возможности оценить значимость, достоверность оценок коэффициентов регрессии. В этом случае выражение (11.2.6) используют для ранжирования факторов.

Рассмотрим геометрическую интерпретацию уравнения регрессии (11.1.5). В случае одной независимой переменной X_1 уравнение линейной регрессии (7.1.1) описывает прямую линию в координатных осях X, Y. Когда же имеются 2 независимых переменных X_1 и X_2, то уравнение (11.1.5) описывает плоскость в системе координат X_1, X_2, Y. Расчетные значения $\hat{Y}_i$ принадлежат этой плоскости, а фактические значения Y_i группируются вблизи этой плоскости по обе ее стороны. В случае одной переменной X близость фактических точек с координатами (X_iY_i) к линии регрессии мы оценивали величиной коэффициента корреляции: чем больше по модулю значение r_{xY}, тем ближе эти точки к линии регрессии. Аналогичную оценку, коэффициент множественной корреляции (КМК) можно использовать и для случая нескольких независимых переменных. КМК можно представить как обычный, парный коэффициент корреляции r_{xY}, если в качестве переменных взять величины фактических Y_i и расчетных $\hat{Y}_i$. Отличие КМК от простого парного r_{xY} в том, что КМК, обозначаемый обычно через R, всегда больше нуля, тогда как $-1 \leqslant r_{xY} \leqslant 1$. Квадрат КМК называют множественным коэффициентом детерминации. Величину $(R')^2$ находят по формуле

$$(R')^2 = \frac{\sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}. \quad (11.3.2)$$

Учитывая ограниченность объема выборки, производят корректировку по формуле

$$R^2 = 1 - \frac{N-1}{N-m-1}[1-(R')^2] \quad . \qquad (11.3.3)$$

Величина R^2 может изменяться в интервале $0 \leqslant R^2 \leqslant 1$. Иногда R^2 обозначают как $R^2_{Y.12}$, указывая, что в данном случае имеется в виду уравнение регрессии $Y = b_0 + b_1X_1 + b_2X_2$. Когда переменные X_1 и X_2 некоррелированы, то $R^2_{Y.12} = r^2_{Yx_1} + r^2_{Yx_2}$.

Проверка достоверности, значимости R^2 производится с помощью статистики

$$F = \frac{R^2(N-m-1)}{m(1-R^2)}, \qquad (11.3.4)$$

которая подчиняется F-распределению Фишера — Снедекора со степенями свободы $f_1 = m$; $f_2 = N-m-1$, где m — число независимых переменных, N — число наблюдений.

11.4. Бета-коэффициенты и ранжирование входных переменных по их влиянию

Заканчивая рассмотрение теоретических основ множественного корреляционно-регрессионного анализа, остановимся на одном очень важном аспекте этого анализа. В разделе 7.1, при обсуждении основных задач проводимого анализа, мы отмечали, что одной из целей регрессионного анализа является ранжирование входных независимых переменных по силе влияния на выходную зависимую переменную. Обратимся к полученному нами ранее выборочному уравнению регрессии (11.1.14). Полученные оценки коэффициентов имеют следующие размерности (с учетом данных табл. 66): b_0(Т/ч), b_1(Т/ч·%), b_2(Т/ч·лет).

При подстановке значений X_1 (%) и X_2 (лет) в уравнение (11.1.14) оцениваемые значения $\hat{Y}_i$ получаем в исходных единицах измерения (Т/ч). Очевидно, что непосредственно по абсолютному значению коэффициентов b_1 и b_2 нельзя сказать, какая из независимых переменных, X_1 или X_2, сильнее влияет, изменяет зависимую переменную Y. Изменив размерности переменных Y, X_1 и

X_2, мы изменяем и величину коэффициентов уравнения регрессии. В том случае, когда для анализа используются безразмерные переменные, или, как их принято называть, стандартизованные переменные:

$$t_Y = \frac{Y - \bar{Y}}{s_Y}; \quad t_{X_1} = \frac{X_1 - \bar{X}_1}{s_{X_1}}; \quad t_{X_2} = \frac{X_2 - \bar{X}_2}{s_{X_2}}, \tag{11.4.1}$$

коэффициенты уравнения регрессии также получаются безразмерными, стандартизованными. В таком стандартизованном уравнении множественной регрессии свободный член $b_0=0$. Соотношение между стандартизованными коэффициентами, которые принято называть «бета-коэффициентами» β_i и размерными коэффициентами b_i, имеет вид

$$\beta_i = \frac{b_i \cdot s_{X_i}}{s_Y}; \quad b_i = \frac{\beta_i \cdot s_Y}{s_{X_i}}. \tag{11.4.2}$$

Поскольку бета-коэффициенты безразмерные, то их уже можно сравнивать непосредственно друг с другом. Стандартизованные коэффициенты множественной регрессии характеризуют скорость изменения среднего значения зависимой переменной по каждой из независимых переменных при постоянных значениях остальных переменных, включенных в анализ. Каждый из бета-коэффициентов показывает, на какую часть стандартного отклонения s_Y изменилось бы среднее значение зависимой переменной Y, если бы значение соответствующей независимой переменной X_i увеличилось на одно стандартное отклонение s_{X_i}, а прочие переменные остались бы без изменения.

Таким образом, зная сравнительную силу влияния каждой из независимых переменных, мы можем проранжировать их по силе влияния на выходную переменную. Программа П 53/34 для регрессионного анализа с двумя независимыми переменными позволяет вычислять стандартные отклонения. Используя значения этих стандартов, получаем бета-коэффициенты

$$\beta_1 = \beta_{x_1} \approx 0{,}937 \text{ и } \beta_2 = \beta_{x_2} = 0{,}086,$$

что свидетельствует о значительно более сильном влиянии уровня механизации работ на производительность труда, нежели среднего возраста работников.

11.5. Полный корреляционно-регрессионный анализ данных с двумя независимыми переменными

Программа П 53/34, приведенная в библиотеке программ, состоит из трех отдельных программ и позволяет провести достаточно полный корреляционно-регрессионный анализ данных с двумя независимыми переменными. Очень удобно использование программ П 53/34 на ПМК «Электроника МК-52», когда три программы последовательно вызываются из энергонезависимой памяти.

Пример. Для демонстрации работы программы П 53/34 используем данные, приведенные в табл. 66.

Решение. Согласно инструкции к П 53/34 вводим программу и начинаем ввод табличных данных: 20 ПВ 32 ПС 33 ПД В/О С/П (результат после останова 1). 24 ПВ 30 ПС 31 ПД В/О С/П (2) и т. д. Вводим последнюю строку: 48 ПВ 76 ПС 41 ПД В/О С/П (14). Ввод таблицы окончен. Вычисляем среднее значение производительности труда по 14 предприятиям: С/П ($35{,}142857=\bar{Y}$). Вычисляем среднее значение коэффициента механизации работ: С/П ($51{,}714285=\bar{X}_1$) и средний возраст работников: С/П ($39=\bar{X}_2$). Далее по п. 9÷11 определяем стандартные отклонения: С/П ($8{,}0751994=s_Y$) С/П ($14{,}39246=s_{X_1}$) С/П ($4{,}38529=s_{X_2}$). Вызываем из регистров памяти и записываем элементы матрицы X^TY: ИП4 ($492=\sum\limits_{i=1}^{14} Y_i$) ИП7 ($26907=\sum\limits_{i=1}^{14} X_{1i}Y_i$) ИП8 ($19384=\sum\limits_{i=1}^{14} X_{2i}Y_i$).

Подготавливаем информационную матрицу Фишера к обращению (п. 13 инструкции): ИП2 (724) П4 ИП3 (546) П7 ИП6 (28533) П8.

Начинаем ввод программы П 53/34—2: В/О F ПРГ (ввод программы) FABT В/О. Содержимое регистров памяти прямого доступа перед началом работы второй программы (п. 15) таково:

Р0=00000000	Р6=28533	РС=76
Р1=14	Р7=546	РД=41
Р2=724	Р8=28533	
Р3=546	Р9=21544	
Р4=724	РА=18138	
Р5=40134	РВ=48	

В регистрах 1—9 хранится информационная матрица, там же будет храниться и обращенная матрица. Подготавливаем операционный стек к работе второй программы: $C_x(0)$ ↑ ↑ ↑ (0). Производим обращение информационной матрицы: 3 С/П ($-1{,}5326748\cdot10^{-1}$). В случае правильного обращения элементы матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$ в регистрах 1—9:

$P7=-1{,}5326748\cdot10^{-1}$; $P8=-5{,}6768753\cdot10^{-4}$;
$P9=4{,}6031328\cdot10^{-3}$; $P4=-2{,}300099\cdot10^{-3}$;
$P5=4{,}2733464\cdot10^{-4}$; $P6=-5{,}0768753\cdot10^{-4}$;

$P1=6{,}1678084$; $P2=-2{,}3001\cdot10^{-3}$; $P3=-1{,}5326748\cdot10^{-1}$. Для контроля эти результаты могут быть извлечены командами ИПN из соответствующих регистров N. По п. 16 вводим в ПМК элементы матрицы $\mathbf{X}^T\mathbf{X}$: 492 ПА 26907 ПВ 19384 ПС. Вычисляем выборочные оценки коэффициентов уравнения регрессии: С/П ($1{,}7361=b_0$) С/П ($5{,}259463\cdot10^{-1}=b_1$) С/П ($1{,}59178\cdot10^{-1}=b_2$).

Вводим программу П 53/34—3: В/О F ПРГ (ввод программы) F АВТ. Вызываем из регистра памяти Р9 элемент обратной матрицы s_{33} и записываем его: ИП9 ($4{,}6031328\cdot10^{-3}$). Далее C_x (0) ПА П9 35,142857 ПО. Начинаем повторный ввод табл. 41 для нахождения оценок по п. 27—39: 20 ↑ 32 ↑ 33 В/О С/П и т. д. После ввода всей таблицы вычисляем $(R')^2$ по п. 27: С/П ($9{,}4484645\cdot10^{-1}=(R')^2$); далее 14 ПО 4,6031328 ВП 3 /—/ П9 С/П ($9{,}6686012\cdot10^{-1}=R$), С/П ($160{,}46308=F_{расч}$), С/П ($4{,}2506293=\hat{\sigma}^2$) $\sqrt{\ }$ ($2{,}0617054=\hat{\sigma}$) С/П ($5{,}1202604=s_{b_0}$) С/П ($3{,}3906478\cdot10^{-1}=t_{b_0}$), С/П ($4{,}2620313\cdot10^{-2}=s_{b_1}$) С/П ($12{,}340273=t_{b_1}$), С/П ($1{,}3987927\cdot10^{-1}=s_{b_2}$) С/П ($1{,}137967=t_{b_2}$). Вычисляем значение стандартизованного бета-коэффициента β_1: ИПС ($5{,}259463\cdot10^{-1}$) $14{,}39246\times(7{,}569661)$ $8{,}0751994\div(9{,}3739617\cdot10^{-1}=\beta_1)$ ИПВ $4{,}38529\times8{,}0751994\div(8{,}6442656\cdot10^{-2}=\beta_2)$.

Вычисления по программе окончены. Теперь проведем анализ полученных результатов. Не повторяя то, что было отмечено ранее, остановимся на проверке значимости полученной зависимости в целом и отдельных ее элементов. Для проверки значимости оценки R^2 сравним полученное значение $F_{расч}=160{,}46$ с табличным $F_{0,05;\,2;\,11}=3{,}98$. Поскольку $F_{расч}>F_{табл}$, то нулевую гипотезу о равенстве $R^2=0$ отклоняем, т. е. полученная

оценка R^2 является достоверной с $P_{дов}=95\%$ (поскольку $F_{расч} \gg F_{табл}$, то фактически $P_{дов}>95\%$). Проверим значимость b_0 и b_1: $t_{0,05;\,11}=1,796$, поэтому на уровне $P_{дов}=95\%$ достоверной можно признать только оценку b_1, поскольку $t_{b_1}=12,34>1,796=t_{0,05;\,11}$.

Поскольку $R^2 \approx 94\%$, то это означает, что такая доля дисперсии производительности труда обусловлена вариациями X_1 и X_2 на исследованных предприятиях.

11.6. Уравнения регрессии с числом входных переменных более двух

При моделировании многомерных наблюдений регрессионными уравнениями нередко приходится иметь дело с набором взаимозависимых переменных, которые фактически все являются выходными переменными. Например, это может быть набор параметров качества прибора либо набор характеристик свойств исходных материалов, которые закоррелированы между собой. При этом возникает необходимость в построении уравнения регрессии для одного из этих признаков как функции всех остальных. Предположим, четыре признака материала измеряются быстро, а пятый признак очень трудно измерить непосредственно, поэтому его нужно оценить по результатам измерения других четырех характеристик. Формально речь идет о таких же уравнениях регрессии, однако в данной ситуации неправомерно будет делить факторы как независимые, входные признаки системы, и зависимые, выходные признаки. Если таких признаков три, то можно воспользоваться программой П 53/34 и получить требуемые уравнения регрессии. Программы П 54/34 и П 55/34 [7] используются в тех случаях, когда таких признаков 4 и 5, т. е. с помощью этих программ можно получить уравнения регрессии с тремя или четырьмя переменными справа.

Коротко рассмотрим реализованный в этих программах метод.

Выше мы уже упоминали, что для стандартизованных значений всех участвующих в регрессии переменных мы имеем уравнение регрессии без свободного чле-

на. При этом в качестве коэффициентов уравнения выступают бета-коэффициенты, т. е.

$$\bar{t}_{1.2,3,\ldots p} = \beta_2 t_2 + \beta_3 t_3 + \ldots + \beta_p t_p, \qquad (11.6.1)$$

где t_2, t_3, ..., t_p — стандартизованные значения переменных X_2, X_3, ..., X_p, $\bar{t}_{1.2,3,\ldots,p}$ — среднее значение стандартизованной переменной X_1, соответствующее заданным значениям переменных t_2, t_3, ..., t_p.

В этом случае система нормальных уравнений, получаемая из выражения (11.1.7), имеет следующий вид:

$$\left.\begin{array}{l} \beta_2 + \beta_3 r_{32} + \beta_4 r_{42} + \ldots + \beta_p r_{p2} = r_{12}, \\ \beta_2 r_{23} + \beta_3 + \beta_4 r_{43} + \ldots + \beta_p r_{p3} = r_{13}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \beta_2 r_{2p} + \beta_3 r_{3p} + \beta_4 r_{4p} + \ldots + \beta_p = r_{1p}, \end{array}\right\} \qquad (11.6.2)$$

где r_{ij} — коэффициенты корреляции между переменными X_i и X_j. Решение системы (11.6.2) дает нам значение бета-коэффициентов, преобразовав которые, используя (11.4.2), получим размерные коэффициенты b_i. Свободный член b_1 уравнения

$$\bar{X}_{1.2,3,\ldots,p} = b_1 + b_2 X_2 + b_3 X_3 + \ldots + b_p X_p \qquad (11.6.3)$$

получаем по формуле

$$b_1 = \bar{X}_1 - b_2 \bar{X}_2 - b_3 \bar{X}_3 - \ldots - b_p \bar{X}_p. \qquad (11.6.4)$$

Итак, для нахождения уравнения регрессии с тремя и четырьмя переменными справа необходимо предварительно с использованием предыдущих программ вычислить следующие величины: средние значения $\bar{X}_i$, стандартные отклонения s_{x_i}, коэффициенты корреляции r_{ij}.

Пример (контрольный)

$r_{23}=0{,}26$; $r_{24}=-0{,}083$; $r_{34}=-0{,}799$; $r_{12}=0{,}8$; $r_{13}=0{,}21$; $r_{14}=-0{,}119$; $\bar{X}_1=4{,}37$; $s_{x_1}=2{,}61$; $\bar{X}_2=5{,}54$; $s_{x_2}=3{,}94$; $\bar{X}_3=0{,}62$; $s_{x_3}=0{,}08$; $\bar{X}_4=0{,}92$; $s_{x_4}=0{,}13$.

Выполнив инструкцию к П 54/34, получим следующие значения β_i и b_i: $\beta_2=8{,}191979\cdot10^{-1}$; $\beta_3=-1{,}2097851\cdot10^{-1}$; $\beta_4=-1{,}4766842\cdot10^{-1}$; $b_2=5{,}426666\cdot10^{-1}$; $b_3=-3{,}9469236$; $b_4=-2{,}9647274$; $b_1=6{,}5382688$.

Коэффициент множественной корреляции R' можно вычислить по формуле

$$R' = \sqrt{\beta_2 r_{12} + \beta_3 r_{13} + \beta_4 r_{14} + \beta_5 r_{15}}. \qquad (11.6.5)$$

ЗАДАЧИ К ГЛ. 11

1[30, с. 113]. Изучалась возможность предсказания по двум тестам производительности труда учеников-клепальщиков. Зависимой переменной в данном случае является производительность при установке заклепок, независимыми переменными — результаты, достигнутые учениками при выполнении теста ловкости рук X_1 и теста ловкости пальцев X_2.

Таблица 67

Число установленных заклепок Y	X_1	X_2	Y	X_1	X_2	Y	X_1	X_2
230	135	107	146	125	78	189	112	80
81	93	67	196	114	89	184	109	85
100	108	81	202	128	84	174	113	75
212	138	93	203	129	80	168	113	87
216	123	81	201	125	99	143	104	69
156	116	86	195	120	86	131	103	65
201	119	86	180	126	92	130	125	84
194	112	96	174	136	95	198	116	76
164	128	80	120	104	82	166	116	86

Используя программу П 53/34, провести множественный корреляционно-регрессионный анализ данных.

2[31, с. 171]. Было решено изучить воздействие климатических условий на заболевание пчел. В одном случае климатические условия измерялись путем непосредственных метеорологических наблюдений, а в другом — с помощью такого вторичного эффекта, как срок начала цветения некоторых летних растений. Данные о наблюдениях содержатся в табл. 68,

где X_1 — процент ульев, подвергшихся заболеванию; X_2 — средняя температура января (по Фаренгейту); X_3 — средняя температура июня (по Фаренгейту); X_4 — дата начала цветения некоторого летнего растения (дни, считая от 1 января).

Таблица 68

№ п/п	X_1	X_2	X_3	X_4	$\widehat{X}_1$
1	49	35	53	200	47,57881
2	40	35	53	212	40,89096
3	41	38	50	211	41,07445
4	46	40	64	212	45,56137
5	52	40	70	203	52,5623
6	59	42	68	194	57,32896
7	53	44	59	194	54,76386
8	61	46	73	188	63,15203
9	55	50	59	196	54,8866
10	64	50	71	190	62,20063

Используя программы П 44/34 и П 54/34, провести множественный линейный регрессионный анализ.

3[28, с. 196]. Исследовалась зависимость коэрцитивной силы электротехнического железа типа «Армко» от величины зерна микроструктуры железа, содержания углерода и серы. Исследования проводились с целью

Таблица 69

Характеристики железа	Коэрцитивная сила X_1	Зернистость X_2	Содержание углерода X_3	Содержание серы X_4
Коэрцитивная сила	1	0,555	0,301	0,372
Зернистость	0,555	1	0,029	0,179
Содержание углерода	0,301	0,029	1	0,334
Содержание серы	0,372	0,179	0,334	1

снижения коэрцитивной силы и улучшения магнитных свойств железа, применяемого в слаботочной аппаратуре. В результате обработки данных по 190 образцам были получены коэффициенты корреляции, приведенные в табл. 69.

Средние значения и стандартные отклонения имели следующие значения:

$\bar{X}_1 = 0{,}929$; $s_{x_1} = 0{,}205$; $\bar{X}_3 = 0{,}015$; $s_{x_3} = 0{,}004$;
$\bar{X}_2 = 3{,}87$; $s_{x_2} = 0{,}893$; $\bar{X}_4 = 0{,}0285$; $s_{x_4} = 0{,}0082$.

Используя программу П 54/34, провести регрессионный анализ данных. Определить величину коэффициента множественной корреляции и проверить его значимость при $\alpha = 5\%$.

РЕШЕНИЯ

1. Ниже приведены результаты обработки данных отдельно по каждой из трех программ. Помимо этого приведено содержание регистров прямого доступа перед вводом в П 53/34—2 и П 53/34—3, т. е. после п. 13 и п. 17 инструкции к П 53/34.

Результаты по П 53/34—1

$\bar{Y} = 172{,}37037$; $\bar{X}_1 = 118{,}14814$; $\bar{X}_2 = 84{,}037037$.

$s_Y = 36{,}538861$; $s_{x_1} = 11{,}002464$; $s_{x_2} = 9{,}488799$.

$$\sum_{i=1}^{27} Y_i = 4654; \quad \sum_{i=1}^{27} X_{1i} Y_i = 556637; \quad \sum_{i=1}^{27} X_{2i} Y_i = 396486.$$

Содержимое регистров памяти после п. 13.

P0 = 00000000;	P5 = 380040;	PA = 836924;
P1 = 27;	P6 = 269820;	PB = 166;
P2 = 3190;	P7 = 2269;	PC = 116;
P3 = 2269;	P8 = 269820;	PД = 86;
P4 = 390;	P9 = 193021;	

Результаты по П 53/34—2

$b_0 = -104{,}02$; $b_1 = 1{,}49844$; $b_2 = 1{,}182246$.

Содержимое регистров памяти после п. 17

P0 = 00000007; P5 = $5{,}4014622 \cdot 10^{-4}$; PA = 0;
P1 = 4,7246954; P6 = $-4{,}0191627 \cdot 10^{-4}$; PB = 1,182246;
P2 = $-3{,}0042 \cdot 10^{-2}$; P7 = $-1{,}3545465 \cdot 10^{-2}$; PC = 1,49844;
P3 = $-1{,}3544648 \cdot 10^{-2}$; P8 = $-4{,}0190935 \cdot 10^{-4}$;
PД = —104,02; P4 = $-3{,}0041419 \cdot 10^{-2}$;
P9 = $7{,}2623107 \cdot 10^{-4}$.

Результаты по П 53/34—3

Выполнив программу, идентифицируйте приведенные ниже числа как оценки тех или иных статистик, полученных при выполнении пп. 18÷39.
23,007875; $6{,}5721997 \cdot 10^{-1}$; $4{,}756351 \cdot 10^{-1}$;
27,539295; 758,41279; —1,7377093; 59,860415;
2,3411587; 1,5930067; $6{,}4004202 \cdot 10^{-1}$;
$7{,}4214751 \cdot 10^{-1}$; $4{,}5120541 \cdot 10^{-1}$; $3{,}0701819 \cdot 10^{-1}$.

Проведите анализ результатов при $\alpha = 0{,}05$.

2. Для контроля результатов по данному заданию приведены две группы чисел, представляющих собой результаты обработки таблицы с данными по программе П 44/34 и П 54/34. Обработав исходные данные, идентифицируйте приведенные числа как оценки тех или иных статистик, получаемые при помощи используемых программ.

I $5{,}8191437 \cdot 10^{-1}$; $-7{,}3333333 \cdot 10^{-1}$; $-6{,}1101009 \cdot 10^{-1}$;
$7{,}9655466 \cdot 10^{-1}$; $8{,}0426952 \cdot 10^{-1}$; $-9{,}363887 \cdot 10^{-1}$;
83,333333; 70; 30; 66; 52; 42; 62; 200.

II 134,29029; $-5{,}573207 \cdot 10^{-1}$; $2{,}0623076 \cdot 10^{-1}$;
$3{,}308412 \cdot 10^{-1}$; $3{,}4071922 \cdot 10^{-1}$; $-6{,}2624256 \cdot 10^{-1}$;
$1{,}3904075 \cdot 10^{-1}$.

На основе полученного уравнения множественной регрессии получены оценки X_1 как функции от X_2, X_3 и X_4. Оценки приведены в таблице. Проверьте эти оценки самостоятельно.

3. Выполнив инструкцию к П 54/34, идентифицируйте приведенные ниже числа как оценки тех или иных статистик:
$5{,}1147043 \cdot 10^{-1}$; $2{,}1666878 \cdot 10^{-1}$; $2{,}0807941 \cdot 10^{-1}$;
11,104275; 5,201985; $1{,}1741482 \cdot 10^{-1}$, $1{,}5978396 \cdot 10^{-1}$.

Используя выражение (11.8.5), определяем значение $R = 6{,}5306119 \cdot 10^{-1}$. Для проверки достоверности этого значения находим F-статистику по (11.3.4): $F = 46{,}106$. Сравнивая это значение с табличным при $\alpha = 5\%$, делаем вывод о достоверности полученного уравнения регрессии.

ГЛАВА 12

ОСНОВЫ КОВАРИАЦИОННОГО АНАЛИЗА

Наука заключается в такой группировке фактов, которая позволяет выводить на основании их общие законы или заключения...

Ч. Дарвин

Выше мы рассмотрели в отдельности методы дисперсионного и регрессионного анализа. Однако нередко данные экспериментальных наблюдений обладают такой структурой, что часть переменных имеет количественный характер и может быть использована для регрессионного анализа, а другая часть имеет качественный характер и может быть использована при проведении дисперсионного анализа. Приведем несколько подобных примеров. Исследуется зависимость урожайности от количества осадков (удобрений) для 4 различных сортов злаковой культуры. Исследуется зависимость среднегодовой зарплаты рабочих от непрерывного стажа работы на предприятии 3 различных министерств (различного профиля, места расположения и т. п.). Исследуется взаимосвязь денежных расходов на удовлетворение материальных потребностей и общей суммы доходов для трех категорий семей: с низким доходом, средним доходом и высоким доходом и т. д.

Ковариационный анализ представляет собой совокупность методов, объединяющих отличительные черты и достоинства как дисперсионного анализа, так и регрессионного анализа. Именно благодаря этому симбиозу ковариационный анализ обладает возможностью выявления более тонкой, глубокой структуры взаимозависимости переменных в экспериментальных наблюдениях.

12.1. Основные цели и возможности ковариационного анализа

Термин «ковариационный анализ» был введен Р. А. Фишером [3, с. 392; 31]. Для объяснения основной идеи

ковариационнго анализа обратимся к рис. 20. На этом рисунке изображены два корреляционных эллипса рассеяния наблюдений, которые отвечают двум группам наблюдений. Проведенные через эти эллипсы пунктирные линии параллельны, т. е. угловые коэффициенты для уравнений регрессии в обеих группах принимаются равными. Обратим внимание на тот факт, что среднее зна-

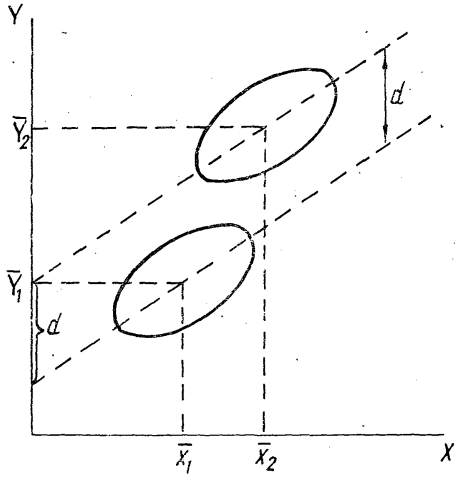

Рис. 20. Расположение корреляционных эллипсов при ковариационном анализе

чение $\bar{X}_2$ для второй группы больше, чем аналогичное среднее значение $\bar{X}_1$ для первой группы. В том случае, если эти средние для обеих групп одинаковы, то оба эллипса будут расположены вертикально один над другим. При этом разность $\Delta Y = \bar{Y}_{.2} - \bar{Y}_{.1}$ будет близка к величине d. Из рисунка ясно, что объединение наблюдений в группу 1 и 2 будет при этом сказываться на величине свободного члена уравнения регрессии.

Для анализа подобных наблюдений первоначально и были развиты методы ковариационного анализа. Переменную X в этом случае обычно называют сопутствующей переменной (она сопутствует исходной группировке наблюдений). Группирующую качественную переменную принято называть индикаторной переменной.

Ковариационный анализ можно продуктивно использовать в следующих целях:

1. Для проверки различий в коэффициентах наклона уравнений регрессии в группах.

2. Для проверки различий в свободных членах уравнений регрессии в группах при равных коэффициентах наклона.

3. Для проверки различий полученных уравнений регрессии в целом, не разделяя их в отдельности на угловой коэффициент и свободный член.

Далее мы рассмотрим особенности достижения перечисленных целей и затем дадим программу для одного случая.

Приведем типичную задачу, которая может быть решена с помощью ковариационного анализа. В табл. 70 приведены данные о трех группах семей, отличающихся между собой средней ежемесячной зарплатой (в рублях) на одного члена семьи — переменная X.

Таблица 70

Низкие доходы		Средние доходы		Высокие доходы	
X	Y	X	Y	X	Y
80	35	150	60	230	95
80	45	210	90	230	120
100	45	200	110	290	125
140	65	160	85	280	140
110	35	190	90	240	130
120	65	180	95	260	125
130	50	180	70	250	110
$\overline{X}_{\cdot 1}=108{,}6$	$\overline{Y}_{\cdot 1}=48{,}6$	$\overline{X}_{\cdot 2}=181{,}4$	$\overline{Y}_{\cdot 2}=85{,}7$	$\overline{X}_{\cdot 3}=254{,}3$	$\overline{Y}_{\cdot 3}=120{,}7$

Первая группа характеризуется низким уровнем доходов на 1 члена семьи, вторая — средним уровнем, третья — высоким. Переменной Y является сумма ежемесячных расходов на удовлетворение материальных потребностей в среднем на одного члена семьи. Нас интересует, зависит ли величина расходов в большей степени от уровня доходов (низкий, средний, высокий) или же от сопутствующей переменной X. Иными словами, бу-

дут ли расходы отличаться между собой для этих групп, если устранить влияние сопутствующей переменной X? Надо сказать, что такая постановка вопроса отнюдь не столь проста, как может показаться на первый взгляд. Ведь уровень доходов в определенной мере формирует психологию и типологию потребления, что, в свою очередь, сказывается и на величине расходов. Отметим, что ответ на этот вопрос как раз и осложняется тем, что средние значения $\bar{X}_{.j}$ увеличиваются при переходе от первой группы ко второй и к третьей. Для устранения влияния сопутствующей переменной X рассмотрим корректирующую модель регрессии

$$Y_{ij}=\mu_{.j}+\beta(X_{ij}-\bar{X}_{..})+\varepsilon_{ij},\ j=1,\ 2,\ \ldots,\ K. \qquad (12.1.1)$$

В такой модели каждое значение Y_{ij} (i — номер наблюдения, строки: j — номер группы, столбца) является суммой среднего значения по столбцу (группе) $\mu_{.j}$, вклада разности $\Delta_{ij}=X_{ij}-\bar{X}_{..}$ (где $\bar{X}_{..}$ — общее среднее значение X по всем наблюдениям) и ошибки ε_{ij}. Из этого становится ясно, что величина углового коэффициента β отражает силу, степень влияния сопутствующей переменной X на Y. Для сравнения средних расходов разных групп скорректируем эти средние расходы в группах следующим образом:

$$\bar{Y}_{\cdot j}^{*}=\bar{Y}_{.j}-\hat{\beta}(\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot}),\quad j=1,2\ldots,K. \qquad (12.1.2)$$

Рассмотрим более внимательно процедуру корректировки с помощью (12.1.2). Для нашей таблицы $\bar{X}_{..}=181{,}4$; $\bar{Y}_{..}=85$. Тогда при $\beta>0$ для первого столбца ($j=1$) разность $\bar{X}_{.1}-\bar{X}_{..}=108{,}6-181{,}4=-72{,}8<0$. В этом случае получаем в результате коррекции увеличение среднего расхода для первой группы, т. е. $\bar{Y}_{.1}^{*}>\bar{Y}_{.1}$. Аналогичным образом для третьей группы получим уменьшение среднего расхода, т. е. $\bar{Y}_{.3}^{*}<\bar{Y}_{.3}$. После проведения такой коррекции можно провести однофакторный дисперсионный анализ и оценить влияние группирующего качественного фактора — уровень доходов. В принципе, в подобной задаче группирующим фактором может быть и любой другой, например, пол, возраст, профессиональная принадлежность и т. д. Важно при этом, что влияние сопутствующей переменной X уже не будет сказываться. Итак, роль регрессионного анализа в ковариационном анализе сводится к оценке величины коэффи-

циента $\hat{\beta}$ при сопутствующей переменной X, а затем уже проводится дисперсионный анализ. Такова общая схема ковариационного анализа.

Прежде чем перейти к оценке величины $\hat{\beta}$, отметим, что реализация изложенной схемы ковариационного анализа имеет смысл тогда, когда генеральные совокупности, из которых взяты изучаемые группы, описываются уравнениями регрессии, которые отличаются только свободными членами. Фактически это означает параллельность линий регрессии (см. рис. 20). Используя программу П 43/34, обработаем каждую из выборок и в результате получим следующие уравнения регрессии:

для группы 1 : $Y=0{,}376\cdot X+7{,}739$; (12.1.3)

для группы 2 : $Y=0{,}532\cdot X-11{,}755$; (12.1.4)

для группы 3 : $Y=0{,}364\cdot X+28{,}051$. (12.1.5)

Далее, обработав все три группы совместно, получим следующее уравнение регрессии:

$$Y=0{,}4865\cdot X-3{,}227. \qquad (12.1.6)$$

Введем следующие условные обозначения, придерживаясь [26, 30]:

$$E_{YX}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\cdot j})\cdot(X_{ij}-\bar{X}_{\cdot j});$$

$$E_{YY}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\cdot j})^2;$$

$$E_{XX}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\cdot j})^2;$$

$$T_{YX}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\cdot\cdot})\cdot(X_{ij}-\bar{X}_{\cdot\cdot});$$

$$T_{YY}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2;$$

$$T_{XX}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\cdot\cdot})^2;$$

$$M_{YX}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot})\cdot(\bar{X}_{j}-\bar{X}_{\cdot\cdot});$$

$$M_{YY} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2;$$

$$M_{XX} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2,$$

где K — число групп (столбцов); n_j — число наблюдений в j-той группе.

Оценка углового коэффициента $\hat{\beta}$ в уравнении (12.1.2) для корректировки групповых средних находится как

$$\hat{\beta} = \frac{E_{YX}}{E_{XX}}.$$

Используя данные таблицы, вычислим $E_{YX}=3907{,}15$ и $E_{XX}=9342{,}87$ и определим $\hat{\beta}$ согласно (12.1.7) — $\hat{\beta}=0{,}4182$. Теперь, зная β, можно оценить и скорректированные значения $\bar{Y}^*_{.j}$ по (12.1.2): $\bar{Y}^*_{.1}=79{,}044662$; $\bar{Y}^*_{.2}=85{,}7$; $\bar{Y}^*_{.3}=90{,}21352$.
Видно, что эти значения имеют уже меньший разброс, нежели исходные $\bar{Y}_{.1}$, $\bar{Y}_{.2}$ и $\bar{Y}_{.3}$. Скорректировав аналогичным образом и остальные наблюдения в каждом столбце, можно будет проводить и дисперсионный анализ, проверяя нулевую гипотезу о равенстве средних значений $\bar{Y}^*_{.j}$, по скорректированным столбцам. Зная $\hat{\beta}$, можно записать внутригрупповые линии регрессии следующим образом:

для группы 1 : $Y=48{,}6+0{,}4182(X-108{,}6)$;

для группы 2 : $Y=85{,}7+0{,}4182(X-181{,}4)$;

для группы 3 : $Y=120{,}7+0{,}4182(X-254{,}3)$.

В том случае, если дисперсионный анализ таких скорректированных данных позволит нам принять нулевую гипотезу

$$H_0 : \mu^*_{.1}=\mu^*_{.2}= \ldots =\mu^*_{.K}=\mu^*_{..} \qquad (12.1.8)$$

против альтернативной гипотезы H_1: не все средние по столбцам равны, то фактически это означает бесполезность принятой классификации по столбцам и тогда вместо уравнения (12.1.1) можно принять следующую модель:

$$Y_{ij}=\mu_{..}+\beta_G(X_{ij}-\bar{X}_{..})+\varepsilon_{ij}. \qquad (12.1.9)$$

Сравним между собой (12.1.1) и (12.1.9). Эти два уравнения отличаются свободными членами ($\mu_{.j}$ и $\mu_{..}$) и угловыми коэффициентами (β и β_G). Оценка β_G находится как

$$\hat{\beta}_G = \frac{T_{YX}}{T_{XX}}. \qquad (12.1.10)$$

В приводимой ниже программе П 56/34 определяются оценки $\hat{\beta}$ и $\hat{\beta}_G$ и проводится проверка следующих нулевых гипотез с помощью дисперсионного анализа:

1) $H_0 : \mu_{.1} = \mu_{.2} = \ldots = \mu_{.K} = \mu_{..}$, (12.1.11)
   $H_1 : \mu_{.1} \neq \mu_{.2} \neq \ldots \neq \mu_{.K} \neq \mu_{..}$,

2) $H_0 : \beta = 0$,
   $H_1 : \beta \neq 0$, (12.1.12)

а также нулевой гипотезы (12.1.8).

Наибольший интерес представляет собой проверка гипотез (12.1.8) и (12.1.12). Принятие нулевой гипотезы (12.1.12) означает, что внутригрупповой коэффициент корреляции равен нулю. Кроме этих гипотез может быть проверена и гипотеза о равенстве β для всех K групп. Кроме одной сопутствующей и индикаторной переменных, могут быть рассмотрены модели и с большим их числом.

12.2. Программа ковариационного анализа данных с одной сопутствующей переменной

Программа П 56/34 вычисляет значение F-критерия Фишера F_1 для проверки нулевой гипотезы о равенстве средних значений $\bar{Y}_{.j}$ (не скорректированных) по столбцам. F-критерий F_2-позволяет провести проверку аналогичной нулевой гипотезы уже для скорректированных средних значений $\bar{Y}^*_{.j}$. Наконец, критерий F_3 позволяет проверить нулевую гипотезу (12.1.12). Отметим также, что величина $\hat{\beta}_G$ представляет собой обычный угловой коэффициент для уравнения регрессии, полученного в результате обработки всех групп в целом.

В том случае, когда нулевая гипотеза (12.1.8) о равенстве скорректированных средних по столбцам отвергается, то, как и ранее при рассмотрении однофакторно-

то дисперсионного анализа, имеет смысл определить те средние (столбцы), которые привели к отклонению нулевой гипотезы. Речь идет о линейных контрастах. Определим линейные контрасты аналогично тому, как это было сделано в главе 6:

$$L = \sum_{j=1}^{k} C_j \mu_{\cdot j}^*, \text{ где } \sum_{j=1}^{k} C_j = 0. \qquad (12.2.1)$$

Используя выражение для коррекции средних по столбцам (12.1.2), получим выражение и для оценки $\hat{L}$:

$$\hat{L} = \sum_{j=1}^{k} C_j \bar{Y}_{\cdot j}^* = \sum_{j=1}^{k} C_j [\bar{Y}_{\cdot j} - \hat{\beta}(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})] \cdot \qquad (12.2.2)$$

В случае нахождения линейного контраста для двух столбцов мы находим оценку разности между скорректированными средними. Например, при $K=3$, $C_1=1$, $C_2=-1$, $C_3=0$ мы находим оценку для разности $\bar{Y}_{\cdot 1}^* - \bar{Y}_{\cdot 2}^*$. Доверительный интервал для $\hat{L}$ строится аналогично (6.2.22) по формуле

$$\hat{L} \pm s_{\hat{L}} \sqrt{(K-1) \cdot F_{\alpha;\, K-1;\, N-K-1}}. \qquad (12.2.3)$$

Значение $s_{\hat{L}}$ находится из выражения

$$S^2_{\hat{L}} = \hat{\sigma}^2 \left\{ \sum_{j=1}^{K} \frac{C_j^2}{n_j} + \frac{\left[\sum_{j=1}^{K} C_j (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot}) \right]^2}{E_{XX}} \right\}, \qquad (12.2.4)$$

$$\text{где } \hat{\sigma}^2 = \frac{1}{N-K-1} \left[E_{YY} - \frac{E^2_{YX}}{E_{XX}} \right]. \qquad (12.2.5)$$

Как и ранее, если доверительный интервал для $\mu_{\cdot 1}^* - \mu_{\cdot m}^*$, т. е. для линейного контраста с $C_l=1$ и $C_m=-1$ и остальными $C_j=0$, покрывает нуль, то принимается нулевая гипотеза о равенстве скорректированных средних значений по l-му и m-му столбцам.

Методы ковариационного анализа тесно соприкасаются с регрессионным анализом, когда для качественных индикаторных переменных вводятся так называемые фиктивные переменные. Достаточно подробно и доступно использование фиктивных переменных и ковариационного анализа рассмотрено в [3, 30, 31].

Пример. Проведем ковариационный анализ данных, представленных в табл. 26. После ввода программы

П 56/34 начинаем ввод $\bar{X}_{..}$ и $\bar{Y}_{..}$: 181,4 ПД 85 П4. Начинаем обработку первого столбца: 108,6 ПО 48,6 П1 80 П9 35 П8 В/О С/П (10281,96) 80 П9 45 П8 В/О С/П (20563,92) и т. д. Обрабатываем последнее, седьмое наблюдение в первом столбце: 140 П9 65 П8 В/О С/П (40413,72). Переходим к обработке второго столбца: 181,4 П0 85,7 П1 150 П9 60 П8 В/О С/П (41399,68) и т. д. ... 210 П9 90 П8 В/О С/П (43099,44). Начинаем обработку третьего столбца: 254,3 П0 120,7 П1 230 П9 95 П8 В/О С/П (45461,4) и т. д. ... 290 П9 125 П8 В/О С/П (83657,16). После обработки всей таблицы значения величин Е и Т следующие: ИП5 ($3907{,}15=E_{YX}$) ИП6 ($9342{,}87=E_{XX}$) ИП7 ($3828{,}58=E_{YY}$) ИПА ($40700=T_{YX}$) ИПВ ($22050=T_{YY}$) ИПС ($83657{,}16=T_{XX}$). Выполняем п. 5, 6 и 7 инструкции: 21 П1 3 П0 С/П (17) ÷ ÷ ПВ ($12{,}656927=F_3$). Вызовем из регистров значения других интересующих нас статистик: ИП3 ($4{,}181959\cdot10^{-1}=\hat{\beta}$) ИП2 ($4{,}8650946\cdot10^{-1}=\hat{\beta}_G$) ИП8 ($42{,}833838=F_1$) ИП9 ($2{,}1084792\cdot10^{-1}=F_2$). Проведем сравнение полученных F-критериев с табличными значениями: $F_1>F_{0{,}05;\,3;\,20}=3{,}1$; $F_2<F_{0{,}05;\,3;\,17}=3{,}2$; $F_3>F_{0{,}05;\,1;\,17}=4{,}5$. Таким образом, отклоняется нулевая гипотеза о равенстве нескорректированных средних значений $\bar{Y}_{.j}$ по столбцам, но принимается аналогичная нулевая гипотеза для скорректированных средних значений $\bar{Y}^*_{.j}$. Отвергается также нулевая гипотеза (12.1.12). Отметим, что полученное значение $\hat{\beta}_G$ равно угловому коэффициенту в уравнении (12.1.6).

ЗАДАЧИ К ГЛАВЕ 12

1[26, с. 297]. В эксперименте над 40 испытуемыми изучалось, насколько близко они могут подойти к устрашающему объекту (живой змее), прежде чем они почувствуют дискомфорт или беспокойство. Всех испытуемых разделили на 4 тренировочных группы, отличающихся по объему тренировки и виду используемого манекена. Одна из групп была контрольной. После тренировки каждого испытуемого подвергали повторному испытанию, измеряя, насколько близко он может подойти к пугающему объекту, не ощущая неудобства или страха. В табл. 71 приведены значения X-расстояния до

тренировки и Y — после тренировки для каждой из 4 групп.

Используя программу П 56/34, провести ковариационный анализ имеющихся данных.

Таблица 71

№ п/п	Тренировочные группы							
	1		2		3		4 (контроль)	
	X	Y	X	Y	X	Y	X	Y
1	25	25	17	11	32	24	10	8
2	13	25	9	9	30	18	29	17
3	10	12	19	16	12	2	7	8
4	25	30	25	17	30	24	17	12
5	10	37	6	1	10	2	8	7
6	17	25	23	12	8	0	30	26
7	9	31	7	4	5	0	5	8
8	18	26	5	3	11	1	29	29
9	27	28	30	26	5	1	5	29
10	17	29	19	20	25	10	13	9

2[26, с. 307]. Исследовалось образование газообразного азота в человеческом организме в естественных условиях. Измеряли величину Y (количество выдыхаемого азота, л) и X (количество вдыхаемого азота, л).

Для 9 человек испытуемых использовали четыре диеты (D_i), отличающиеся содержанием белков. Данные об этом опыте приведены в табл. 72.

Таблица 72

№ п/п	Диета							
	Безбелковая (D_1)		23% белков (D_2)		32% белков (D_3)		67% белков (D_4)	
	Y	X	Y	X	Y	X	Y	X
1	4,079	4,158	4,368	4,332	4,169	4,102	4,928	4,829
2	4,859	4,877	5,668	5,617	5,709	5,582	5,608	5,400
3	3,540	3,576	3,752	3,720	4,416	4,339	4,940	4,799
4	5,047	5,078	5,848	5,797	5,666	5,585	5,291	5,167
5	3,298	3,315	3,802	3,773	4,123	4,049	4,674	4,565
6	4,679	4,702	4,844	4,800	5,059	4,987	5,038	4,933
7	2,870	2,901	3,578	3,539	4,403	4,322	4,905	4,762
8	4,648	4,718	5,393	5,317	4,496	4,383	5,208	5,080
9	3,847	3,880	4,374	4,343	4,688	4,623	4,806	4,709
Среднее:	4,0963	4,1339	4,6252	4,5809	4,7477	4,6636	5,0442	4,9160

С помощью программы П 56/34 провести ковариационный анализ приведенных данных при $\alpha=5\%$.

3[26, с. 311]. Четыре лекарства, снижающих артериальное давление, сравниваются методом ковариационного анализа. Откликом (исследуемой переменной) служит систолическое давление у подопытных животных после курса лечения, а сопутствующей переменной — систолическое давление до лечения. Данные приведены в следующей табл. 73.

Таблица 73

Лекарство	Систолическое давление		Лекарство	Систолическое давление	
	до лечения X	после лечения Y		до лечения X	после лечения Y
	194	157		172	136
1	162	136	3	196	182
	183	145		158	134
	180	153			
	154	124		158	124
2	184	123	4	165	124
	173	143		186	132
	170	136		182	133

Используя программу П 56/34, проведите ковариационный анализ этих данных и сформулируйте полученные выводы относительно исследуемых лекарств. Значение $\alpha=5\%$.

4. Используя числовые данные задач 2 и 3 в качестве основы, составьте аналогичные условия задач по собственной тематике исследований. Дайте содержательную интерпретацию полученным при этом результатам решения.

РЕШЕНИЯ

1. Вначале вычисляем средние значения переменных X и Y в отдельности по группам и в целом по всем 40 испытуемым:

$\bar{X}_{.1}=17{,}1$; $\bar{X}_{.2}=16{,}0$; $\bar{X}_{.3}=16{,}8$; $\bar{X}_{.4}=15{,}3$

$\bar{Y}_{.1}=26{,}8$; $\bar{Y}_{.2}=11{,}9$; $\bar{Y}_{.3}=8{,}2$; $\bar{Y}_{.4}=15{,}3$

$\bar{X}_{..}=16{,}3$; $\bar{Y}_{..}=15{,}5$.

После ввода программы П 56/34 начинаем ввод исходных данных: 16,3 ПД 15,5 П4 17,1 П0 26,8 П1. Начинаем ввод данных первого столбца: 25 П9 25 П8 В/О С/П (75,69) и т. д. Обратим внимание на то, что в случае $X_{ij}=Y_{ij}$ необходимо производить их набор и ввод в регистры 9 и 8 отдельно, как, например, это сделано выше: 25 П9 25 П8, это объясняется построением программы. После окончания вычислений получим следующие результаты: $\hat{\beta}=6{,}4268592\cdot10^{-1}$; $\hat{\beta}_G=6{,}5960381\cdot10^{-1}$; $F_1=8{,}8501253$; $F_2=16{,}443695$; $F_3=34{,}708519$.

Предлагается самостоятельно применить метод линейных контрастов.

2. Предварительно вычисляем средние значения $\bar{Y}_{..}$ и $\bar{X}_{..}$ в целом на 36 наблюдений: $\bar{X}_{..}=4{,}5736$; $\bar{Y}_{..}=4{,}62835$. После выполнения п. 1—4 к П 56/34 получаем следующие значения E_{ij} и T_{ij} (идентифицируйте их, обработав таблицу самостоятельно): 16,616272; 18,289045; 13,892941; 13,747712; 17,37629; 14,056904. Выполняя следующие пункты инструкции, получаем интересующие нас статистики, которые также идентифицируйте самостоятельно: 60,539599; 3,2114352; 25302,645; 1,0105638; 1,0457393. Закончив идентификацию этих статистик, ответьте на следующие вопросы:

1. Можно ли принять нулевую гипотезу о равенстве средних значений количества выдыхаемого азота $\bar{Y}_{.j}$ для всех четырех диет? Если нет, то как применить линейные контрасты?

2. Можно ли принять аналогичную нулевую гипотезу к скорректированным средним значениям?

3. Можно ли принять нулевую гипотезу о равенстве нулю углового коэффициента β (12.1.12)?

3. Предварительно определяем $\bar{X}_{..}$, $\bar{Y}_{..}$, $\bar{X}_{.j}$ и $\bar{Y}_{.j}$: $\bar{X}_{..}=174{,}46666$; $\bar{Y}_{..}=138{,}8$; $\bar{X}_{.1}=179{,}75$; $\bar{Y}_{.1}=147{,}75$; $\bar{X}_{.2}=170{,}25$; $\bar{Y}_{.2}=131{,}5$; $\bar{X}_{.3}=175{,}33333$; $\bar{Y}_{.3}=150{,}66666$; $\bar{X}_{.4}=172{,}75$; $\bar{Y}_{.4}=128{,}25$.

Выполнив полностью инструкцию к П 56/34, имеем следующие результаты: $\hat{\beta}=6{,}805867\cdot10^{-1}$; $\hat{\beta}_G=7{,}9489124\cdot10^{-1}$; $F_1=2{,}4616411$; $F_2=2{,}8750679$; $F_3=10{,}124351$.

Сравнив F_i с соответствующими табличными значениями, можно ли принять следующие утверждения:

1. Для всех четырех лекарств средние значения систолического давления крови у пациентов после лечения равны с $P_{дов}=95\%$.

2. Для использованных лекарств нет зависимости систолического давления после лечения от аналогичного давления до лечения.

ГЛАВА 13

ПРОВЕРКА ГИПОТЕЗ О ВЕКТОРАХ СРЕДНИХ

Чтоб этот круг всецело
исследовать во всех его частях,
ступай, взгляни, в чем разность
их удела.

А. Данте

Очень часто в практике исследований необходимо производить проверку гипотез относительно вектора средних. Приведем несколько примеров, иллюстрирующих ситуации, в которых возникает возможность и необходимость в применении таких гипотез.

1. Антрополог исследует останки людей, живших несколько тысячелетий тому назад. При изучении 25 черепов получены средние значения двух коррелированных между собой размеров черепа. Эти два значения обозначим через вектор средних $\bar{X}^{\mathrm{T}}=[\bar{X}_1;\ \bar{X}_2]$. Для идентификации принадлежности этих людей одной из предполагаемых групп требуется провести сравнение вектора средних $\bar{X}$ с заданным вектором μ_0, который характерен для черепов людей предполагаемой группы.

2. Разработан технологический процесс производства полупроводникового прибора, основные параметры которого (мощность генерации P и частота генерации f) коррелированы между собой. Известны номинальные значения частоты f_0 и мощности P_0, которые образуют заданный вектор $\mu_0^{\mathrm{T}}=[f_0;\ P_0]$. Из выпускаемых приборов взята контрольная выборка объемом в 100 приборов. Получены средние значения $\bar{f}$ и $\bar{P}$, которые образуют вектор средних $\bar{X}^{\mathrm{T}}=[\bar{f};\ \bar{P}]$. Необходимо ответить на вопрос: равен ли вектор генеральных средних исследуемой генеральной совокупности μ заданному вектору μ_0? Иными словами, обеспечивает ли используемый техпроцесс получение приборов с заданными номинальными значениями, но с учетом закоррелированности f и P?

3. По двум отличающимся технологиям производят

алюминиевый сплав. По результатам исследования двух выборок алюминиевого сплава по каждой из технологий получены векторы средних. Эти векторы характеризуют предел выносливости и предел прочности, которые закоррелированы между собой. Требуется произвести сравнение этих двух векторов средних и ответить на вопрс: отличаются ли между собой векторы средних этих двух генеральных совокупностей, или нет? Основное отличие процедуры выполнения подобной проверки от проверки гипотез о равенстве одномерных средних (см. главу 4) в следующем. Когда мы сравниваем два одномерных средних значения двух выборок, то фактически строим для каждого среднего доверительный интервал (одномерный) и если эти интервалы пересекаются, накладываются хотя бы частично, друг на друга, то мы делаем заключение о том, что средние данных генеральных совокупностей статистически неразличимы. Иными словами, разность между этими средними статистически незначимо отличается от нуля. Для этой цели мы используем либо t-распределение Стьюдента, либо нормальное стандартное распределение. Когда же мы вынуждены иметь дело с двумя или большим числом закоррелированных переменных, то по выборке мы определяем уже вектор средних $\bar{X}^{т} = [\bar{X}_1 \bar{X}_2 \ldots \bar{X}_к]$. В этом случае мы уже не можем построить отдельно для каждой переменной доверительный интервал, а строим совместную доверительную область — двухмерный эллипс либо трех- или многомерный эллипсоид.

13.1. Критерий T^2 — Хотеллинга

Для проверки подобных многомерных статистических гипотез вместо выборочной t-статистики используется ее многомерный аналог, так называемый критерий Хотеллинга T^2. Теория статистики T^2 была разработана статистиком Гарольдом Хотеллингом. Первая статья Хотеллинга, посвященная T^2, была опубликована в «Анналах математической статистики» в 1931 г., затем последовало еще несколько его статей, опубликованных в психологических журналах. Статистика T^2 является одной из основных, используемых в многомерном статистическом анализе [32]. Напомним, что выборочная t-ста-

тистика при проверке, например, равенства выборочного одномерного среднего $\bar{X}$ заданному значению (генеральному среднему μ_0), записывается в виде

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}. \qquad (13.1.1)$$

Для демонстрации аналогии между критерием T^2 Хотеллинга и t-критерием Стьюдента возведем обе части (13.1.1) в квадрат

$$t^2 = n(s^2)^{-1} \cdot (X - \mu_0)^2 \qquad (13.1.2)$$

и запишем это выражение иначе

$$t^2 = n(X - \mu_0) \cdot (s^2)^{-1} \cdot (X - \mu_0). \qquad (13.1.3)$$

Аналогичный вид имеет выражение и для критерия T^2 Хотеллинга

$$T^2 = n(\mathbf{X} - \mu_0)^T \cdot \mathbf{S}^{-1} \cdot (\mathbf{X} - \mu_0). \qquad (13.1.4)$$

В выражении (13.1.4) T^2 и n — скаляры; матрица $\mathbf{S}$ — ковариационная матрица

$$\mathbf{S} = \frac{1}{n-1} \cdot \mathbf{X}^T\mathbf{X}, \qquad (13.1.5)$$

где $\mathbf{X} = x - \mu_0$. В одномерном случае матрица $\mathbf{S}$ приводится к s^2. Обратим внимание на то, что в этих выражениях и далее векторы и матрицы обозначаются полужирным шрифтом; T — символ транспонирования.

Известно, что t-распределение и F-распределение связаны между собой следующим соотношением:

$$(t_{\alpha/2;\ f_2})^2 = F_{\alpha;\ f_1;\ f_2}, \qquad (13.1.6)$$

если $f_1 = 1$.

Аналогичное соотношение существует между T^2 и F-распределением:

$$T^2_{\alpha;\ K;\ n-K} = \frac{K \cdot (n-1)}{n-K} \cdot F_{\alpha;\ K;\ n-K}. \qquad (13.1.7)$$

Легко заметить, что при числе переменных $K = 1$ (13.1.7) переходит в (13.1.6), т. е.

$$(t_{\alpha/2;\ n-1})^2 = T^2_{\alpha;\ 1;\ n-1} = F_{\alpha;\ 1;\ n-1}, \qquad (13.1.8)$$

С помощью этого соотношения для выборочной статистики T^2 определяют критическое значение $T^2_{кр}$, ис-

пользуя таблицы F-распределения Фишера — Снедекора.

Далее мы рассмотрим методы проверки гипотезы о равенстве вектора выборочных средних заданному вектору и о равенстве двух векторов выборочных средних, а также программные реализации этих методов для случая двух переменных.

13.2. Проверка статистической гипотезы о равенстве вектора выборочных средних заданному вектору

Рассмотрим последовательность проверки гипотезы на следующем примере. Исследовалась популяция тихоокеанского морского беспозвоночного. Ранее этот вид беспозвоночных был обнаружен в Индийском океане и тщательно изучен. Основные морфологические признаки этого вида (длина и ширина тела) равны 44 мм и 3,1 мм.

Таблица 74

U_i	3,0	3,2	3,7	3,3	3,5	3,2	3,2	3,3	3,4	3,3	3,2	3,3
X_1	—0,3	—0,1	0,4	0	0,2	—0,1	—0,1	0	0,1	0	—0,1	0
f_i	45	42	37	41	40	44	43	42	44	40	45	41
X_2	3	0	—5	—1	—2	2	1	0	2	—2	3	—1

В табл. 74 приведены размеры 12 особей этого вида тихоокеанской популяции.

По данной выборке среднее значение ширины $\bar{U}=3{,}3$ мм, а средняя длина $\bar{f}=42$ мм. В табл. 74 через U_i и f_i обозначены измеренные значения ширины и длины, а через X_{1i} и X_{2i} — эти же переменные в отклонениях от средних значений (центрированные переменные), т. е. $X_{1i}=U_i-\bar{U}$ и т. д. Таким образом, имеется заданный вектор $\mu^T_0=[3{,}1;\ 44]$ и вектор выборочных средних $\bar{X}^T=[3{,}3;\ 42]$. При этом мы предполагаем, что наш вектор выборочных средних $\bar{X}$ представляет собой оценку генерального вектора $\mu^T=[\mu_1\ \mu_2]$. Представляет интерес сравнение вектора μ с вектором μ_0. Итак, сформулируем нулевую и альтернативную гипотезы:

$$H_0: \mu=\mu_0; \qquad H_1: \mu\neq\mu_0. \qquad (13.2.1)$$

Принятие нулевой гипотезы в данном примере означало бы, что исследуемый вид имеет одинаковый морфологический критерий независимо от ареала обитания.

В том случае, когда нулевая гипотеза H_0 верна, критерий Хотеллинга T^2 связан с F-распределением по (13.1.7). Для вычисления T^2 по (13.1.4) сформулируем гипотезу H_0 более конкретно:

$$H_0 : [\mu_1 \quad \mu_2] = [3{,}1 \quad 44], \tag{13.2.2}$$

$$H_1 : [\mu_1 \quad \mu_2] = [3{,}1 \quad 44].$$

Для вычисления выборочной T^2-статистики по (13.1.4) нам потребуются следующие векторы и матрицы:

$$[X-\mu_0]^T = [3{,}3 \;\; 42] - [3{,}1 \;\; 44] = [0{,}2 \;\; -2]; \tag{13.2.3}$$

$$X^T X \begin{bmatrix} -0{,}3 & -0{,}1 & \dots & 0 \\ 3 & 0 & \dots & -1 \end{bmatrix} \cdot \begin{bmatrix} -0{,}3 & 3 \\ -0{,}1 & 0 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & -1 \end{bmatrix} =$$

$$\begin{bmatrix} \sum\limits_{i=1}^{12} X^2_{1i} & \sum\limits_{i=1}^{12} X_{1i} X_{2i} \\ \sum\limits_{i=1}^{12} X_{1i} X_{2i} & \sum\limits_{i=1}^{12} X^2_{2i} \end{bmatrix} = \begin{bmatrix} 0{,}34 & -3{,}7 \\ -3{,}7 & 62 \end{bmatrix}; \tag{13.2.4}$$

$$S = \frac{1}{n-1} X^T X = \frac{1}{11} \cdot \begin{bmatrix} 0{,}34 & -3{,}7 \\ -3{,}7 & 62 \end{bmatrix} =$$

$$= \begin{bmatrix} 3{,}090909 \cdot 10^{-2} & -3{,}3636363 \cdot 10^{-1} \\ -3{,}3636363 \cdot 10^{-1} & 5{,}636363 \end{bmatrix}, \tag{13.2.5}$$

$$S^{-1} = \begin{bmatrix} 92{,}28688 & 5{,}507443 \\ 5{,}507443 & 5{,}060893 \cdot 10^{-1} \end{bmatrix}. \tag{13.2.6}$$

Согласно (13.1.4) вычисляем значение статистики T^2:

$$T^2 = 12 \cdot [0{,}2 \;\; -2] \cdot \begin{bmatrix} 92{,}28688 & 5{,}507443 \\ 5{,}507443 & 5{,}060893 \cdot 10^{-1} \end{bmatrix} \times$$

$$\times \begin{bmatrix} 0{,}2 \\ -2 \end{bmatrix} = 15{,}718537. \tag{13.2.7}$$

Критическое значение T^2 определим по (13.1.7):

$$T^2_{0,05;\,2;\,10} = \frac{2 \cdot (12-1)}{10} \cdot 4{,}10 = 9{,}02, \qquad (13.2.8)$$

т. к. $F_{0,05;\,2;\,10} = 4{,}10$.

Поскольку выборочное значение T^2 превышает критическое значение, то нулевую гипотезу с доверительной вероятностью $P_{дов} = 95\%$ отвергаем. Итак, сравнение вектора выборочных средних с заданным вектором не позволяет утверждать, что обе популяции имеют идентичные морфологические критерии.

Рассмотрим процедуру построения совместной доверительной области — корреляционного эллипса, для вектора генеральных средних μ, оценкой которого является вектор выборочных средних $\bar{X}$. Для совместно рассматриваемых K средних доверительная область будет заключена внутри и на поверхности K-мерного эллипсоида, поверхность которого определена уравнением

$$(\bar{X} - \mu)^T - S^{-1} \cdot (\bar{X} - \mu) = \frac{K(n-1)}{n(n-K)} \cdot F_{\alpha;\,K:\,n-K}. \qquad (13.2.9)$$

Центр этого эллипсоида определен вектором средних $\bar{X}$. Несложно заметить, что для одной переменной, т. е. $K=1$, выражение (13.2.9) превращается в уравнение

$$\frac{(X-\mu)^2}{s^2} = \frac{1 \cdot (n-1)}{n \cdot (n-1)} \cdot F_{\alpha;\,1;\,n-1}. \qquad (13.2.10)$$

Извлекая квадратный корень из обеих частей (13.2.10), с учетом (13.1.8) получаем

$$\frac{\bar{X} - \mu}{s} = \pm \frac{1}{\sqrt{n}} \cdot t_{\alpha/2;\,n-1}, \qquad (13.2.11)$$

т. е. мы получили уже известный нам из главы 3 результат. Это выражение определяет доверительный интервал для одномерного генерального среднего μ по его оценке $\bar{X}$.

При $K=2$ построение доверительной области для генерального вектора μ можно осуществить с использованием рассмотренной выше программы П 52/34. При использовании этой программы изменение касается только

величины χ^2, вносимой в регистр 8. В данном случае вместо χ^2 в регистр 8 вносится значение выражения

$$\frac{2(n-1)}{n\cdot(n-2)}\cdot F_{\alpha;\,2;\,n-2}=\frac{T^2_{крит}}{n}. \qquad (13.2.12)$$

Для использования программы П 52/34 нам необходимо знать коэффициент корреляции r_{12} между двумя переменными. Имея матрицу S, это несложно сделать. Напомним, что

$$r_{12}=\frac{s_{12}}{\sqrt{s_{11}}\cdot\sqrt{s_{22}}}. \qquad (13.2.13)$$

Подставив в (13.2.13) числовые значения нашего примера, получим

$$r_{12}=\frac{-3{,}3636363\cdot10^{-1}}{\sqrt{3{,}090909\cdot10^{-2}}\cdot\sqrt{5{,}636363}}=-8{,}0587266\cdot10^{-1}.$$

Приняв по вертикальной оси f масштаб 1 мм/10 мм, т. е. $m_f=0{,}1$, и значение $f_{min}=40$ мм, которое заносим при наборе программы П 52/34 с 53-го по 59-й адрес (40,0), а также значение выражения (13.2.12), равное $7{,}5166666\cdot10^{-1}$ и занесенное в регистр 8, получаем таб-

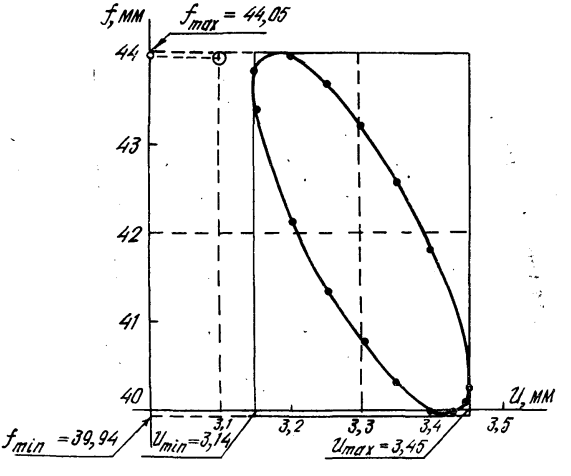

Рис. 21. Доверительная область для вектора μ и доверительные интервалы для линейных контрастов

лицу данных для построения доверительной области. Используя эти данные, построим корреляционный эллипс (рис. 21). На этом же рисунке нанесем точку ⊙, отвечающую базовому вектору μ_0. Видно, что точка ⊙ находится вне доверительной области для μ, что и приводит к отклонению нулевой гипотезы H_0 (табл. 75).

Таблица 75

U, мм	3,15	3,20	3,25	3,30
$f_{верх}$, мм	43,848871	44,008002	43,695389	43,218706
$l_{верх}$, мм	38,48871	40,08002	36,95389	32,18706
$f_{нижн}$, мм	43,415835	42,168468	41,392846	40,781294
$l_{нижн}$, мм	34,15835	21,68468	13,92846	7,81204
U, мм	3,35	3,40	3,45	
$f_{верх}$, мм	42,607154	41,831532	40,584165	
$l_{верх}$, мм	26,07154	18,31532	5,84165	
$f_{нижн}$, мм	40,304611	39,991998	40,151129	
$l_{нижн}$, мм	3,04611	$-8{,}002\cdot10^{-2}$	1,51129	

Аналогично тому, как мы в главе 6 использовали метод линейных контрастов для того, чтобы «заглянуть за кулисы» при отклонении нулевой гипотезы, можно организовать линейные контрасты и получить значения выборочных статистик T^2 для переменной U и для переменной f. Не описывая подробно процедуру построения линейных контрастов и нахождения T^2, укажем лишь конечный результат: $T^2_U = 15{,}529412$, а $T^2_f = 8{,}516129$. Поскольку значение T^2_f меньше, чем критическое значение $T^2_{кр} = 9{,}02$, то делаем вывод о том, что изменение длины животных не является причиной отклонения нулевой гипотезы H_0. И напротив, поскольку T^2_U больше критического значения $T^2_{кр}$, то делаем вывод о том, что причиной отклонения нулевой гипотезы H_0 послужило именно отклонение среднего значения ширины тела. В том случае, когда оба линейных контраста дают значения T^2_i больше критического значения $T^2_{кр}$, это означает, что произошло смещение вектора по обеим составляющим.

Для линейных контрастов также можно найти доверительные интервалы [30]. Не рассматривая подробно процедуру нахождения таких интервалов, укажем их

значения. Для ширины тела U имеем интервал 3,1475752÷3,4524248 мм, для длины тела f : 39,941684÷÷44,058316 мм. Видно, что значение длины $f_0=44$ мм попадает в этот интервал, а значение ширины тела $U_0=3{,}1$ мм находится вне этого интервала. Использование подобных доверительных интервалов по своему смыслу аналогично проверке нулевой гипотезы с использованием линейных контрастов.

На рис. 21 доверительный эллипс обведен прямоугольной рамкой. Линии, проходящие через стороны этого прямоугольника, пересекают оси в точках с координатами, равными границам соответствующих доверительных интервалов для линейных контрастов.

В том случае, когда точка ⊙ лежит вне прямоугольника, нулевая гипотеза отвергается и при использовании доверительного эллипса, и при использовании доверительных интервалов. Когда же точка ⊙ находится вне эллипса, но внутри прямоугольника, то нулевая гипотеза будет отвергнута при использовании доверительного эллипса и принята при использовании контрастов. Это ограничивает возможность анализа причин отклонения H_0 «изнутри». В таких ситуациях необходимо реализовывать более сложные линейные комбинации.

Программа П 57/34 для проверки гипотезы о равенстве вектора средних $\overline{X}$ заданному вектору μ_0 с помощью критерия Хотеллинга T^2 позволяет вычислять достаточно полный набор величин, характеризующих гипотезу. Программа вычисляет значение T^2 по (13.1.4), находит критическое значение статистики $T^2_{кр}$, затем производит проверку нулевой гипотезы с помощью линейных контрастов, определяя значения T^2_1 и T^2_2, и, наконец, находит доверительные границы для этих линейных контрастов. Для работы программы необходимо знание основных характеристик двухмерной системы переменных, т. е. коэффициент корреляции r_{12}, значения векторов $\overline{X}$ и μ_0, значения выборочных дисперсий $S^2_{x_1}$ и $S^2_{x_2}$, объем выборки n, значение $F_{\alpha;\,2;\,n-2}$ и значения средних $\overline{X}^T=[X_1\ X_2]$ и $\mu_0=[\mu_{01}\ \mu_{02}]$.

Пример. Проиллюстрируем работу программы П 57/34 на примере данных табл. 74, приняв $U=X_1$, $f=X_2$.

Решение. Исходные данные вносим в ПМК:

1) объем выборки n : 12 П8;

2) коэффициент корреляции r_{12}: 8,0587266 /—/ ВП 1 /—/ П7;

3) $\Delta X_1 = X_1 - \mu_{01} = 3{,}3 - 3{,}1 = 0{,}2 : 0{,}2$ П6;

4) $\Delta X_2 = X_2 - \mu_{02} = 42 - 44 = -2 : 2$ /—/ П5;

5) дисперсия $s^2_{x_1} = 3{,}090909 \cdot 10^{-2} : 3{,}090909$ ВП 2 ПВ /—/ П1;

6) дисперсия $s^2_{x_2} = 5{,}6363636 : 5{,}636363$ ПА /—/ П3;

7) значение $F_{0{,}05;\ 2;\ 10} = 4{,}1 : 4{,}1$ П4.

По окончании ввода данных вычисляем значение выборочной статистики T^2 : В/О С/П. Результат $T^2 = 15{,}718537$. Далее вычисляем критическое значение статистики T^2 : С/П ($9{,}02 = T^2_{кр}$). Поскольку $T^2 > T^2_{кр}$, то нулевая гипотеза отвергается. Проводим проверку нулевой гипотезы с помощью линейных контрастов: С/П ($8{,}516129 = T^2_{x_2}$), т. е. изменение средней длины тела беспозвоночного не является причиной отклонения нулевой гипотезы. Далее С/П ($15{,}529412 = T^2_{x_1}$), т. е. причина отклонения H_0 в значимом изменении ширины тела. Затем находим верхнюю и нижнюю границу для длины тела: С/П ($44{,}058316 = f_{верх}$) С/П ($39{,}941684 = f_{нижн}$). Аналогично находим границы доверительного интервала для ширины тела: С/П ($3{,}4524248 = U_{верх}$) С/П ($3{,}1475752 = U_{нижн}$).

13.3. Проверка статистической гипотезы о равенстве двух векторов выборочных средних

Рассмотрим следующий пример. Исследуется популяция морских беспанцирных моллюсков. Были изучены 15 особей с глубины 1,5 км и 18 особей с глубины 2,0 км. В результате исследования были определены средние значения морфологических признаков — количество щетинок, покрывающих тело моллюска $\bar{P}$ и длина его тела $\bar{f}$ для обеих выборок. Учитывая, что длина тела и количество щетинок на теле являются коррелированными признаками, необходимо провести проверку гипотезы о равенстве двух векторов выборочных средних.

По первой выборке имеем следующие данные: $\bar{P}_1 = 180$; $\bar{f}_1 = 23$ мм; $s^2_p = 225$; $s^2_f = 10{,}24$ мм^2; $r = -0{,}3$; $n_1 = 15$. Данные по второй выборке: $\bar{P}_2 = 160$; $\bar{f}_2 = 26$ мм; $s^2_p = 400$; $s^2_f = 16$ мм^2; $r = -0{,}4$; $n_2 = 18$.

Рассмотрим последовательность проверки гипотезы о равенстве двух векторов средних $\overline{X}_1$ и $\overline{X}_2$, которые получены на основе выборок из двух многомерных нормальных совокупностей $N_K(\mu_1, \Sigma)$ и $N_K(\mu_2, \Sigma)$. Обе совокупности, из которых взяты выборки, имеют одну и ту же ковариационную матрицу Σ. Однако в силу случайного выбора элементов из двух генеральных совокупностей мы можем получить векторы средних $\overline{X}_1$ и $\overline{X}_2$, отличающиеся друг от друга. Некоторые отличия могут наблюдаться и в выборочных ковариационных матрицах S_1 и S_2. На основе полученных выборок мы хотим проверить нулевую гипотезу

$$H_0 : \mu_1 = \mu_2$$

против (13.3.1)

$$H_1 : \mu_1 \neq \mu_2$$

Эти же гипотезы можно сформулировать иначе:

$$H_0 : \Delta\mu = 0, \qquad (13.3.2)$$

$$H_1 : \Delta\mu \neq 0.$$

Подобная проверка позволит нам установить, все ли элементы двух векторов генеральных средних одновременно равны, или же часть элементов отличается друг от друга.

Аналогично тому, как мы делали в п. 4.3 при сравнении двух выборочных одномерных средних (выражение 4.3.6), вначале, используя две имеющиеся оценки S_1 и S_2, получим объединенную ковариационную матрицу

$$S_* = \frac{[X_1^T X + X_2^T X_2]}{n_1 + n_2 - 2}, \qquad (13.3.3)$$

где X_1 — это $n_1 \times K$ матрица наблюдений в виде отклонений от среднего ($X_{ik} = x_{ik} - \bar{x}_k$) по выборке из совокупности $N_k(\mu_1, \Sigma)$, а X_2 — это $n_2 \times K$ аналогичная матрица наблюдений из $N_k(\mu_2, \Sigma)$. Как и при сравнении одномерных средних число наблюдений по обеим выборкам может быть разным, но обязательно должно быть равным число переменных.

Действуя по аналогии с (13.1.2), возводим в квадрат выражение для t-статистики (4.3.6), при этом получаем

$$t^2 = \frac{n_1 \cdot n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2) \cdot (S_{**})^{-1} (\bar{X}_1 - \bar{X}_2), \qquad (13.3.4)$$

Далее, заменив скаляры векторами и матрицами, получим выражение для статистики Хотеллинга T^2:

$$T^2 = \frac{n_1 \cdot n_2}{n_1 + n_2} (\overline{X}_1 - \overline{X}_2) \cdot S^{-1}_* \cdot (\overline{X}_1 - \overline{X}_2). \quad (13.3.5)$$

При верной нулевой гипотезе критическое значение статистики T^2 по (13.3.5) связано с F-распределением следующим соотношением:

$$T^2_{\alpha;\, K;\, n_1+n_2-K-1} = \frac{(n_1 + n_2 - 2) \cdot K}{n_1 + n_2 - K - 1} \times$$

$$\times F_{\alpha;\, K; n_1+n_2-K-1}. \quad (13.3.6)$$

В том случае, когда значение T^2 по (13.3.5) больше критического значения T^2 по (13.3.6), нулевая гипотеза отклоняется при уровне значимости α, и наоборот, при $T^2 \leqslant T^2_{кр}$ гипотеза H_0 принимается.

Приводимая в библиотеке программа П 58/34 вычисляет элементы объединенной ковариационной матрицы, обращает эту матрицу и производит умножение трех матриц согласно (13.3.5) и вычисляет значение T^2. В качестве исходных данных для работы этой программы, производящей вычисления различных величин, требующихся при сравнении двухмерных векторов средних, используются те же данные, что и в предыдущей программе П 57/34, но по каждой выборке отдельно.

Так же как и при сравнении вектора средних с заданным вектором, в нашей задаче можно построить доверительную область — многомерный эллипсоид. Такую $100 \cdot (1-\alpha)$ %-ную доверительную область для разности $\Delta\mu = \mu_1 - \mu_2$ можно построить с помощью разности $\Delta\overline{\mathbf{X}} = \mathbf{X}_1 - \mathbf{X}_2$. Уравнение поверхности многомерного эллипсоида имеет вид

$$(\Delta\overline{\mathbf{X}} - \Delta\mu)^T \cdot s_*^{-1} \cdot (\Delta\mathbf{X} - \Delta\mu) = \frac{n_1 + n_2}{n_1 \cdot n_2} \times$$

$$\times \frac{(n_1 + n_2 - 2) \cdot K}{n_1 + n_2 - K - 1} \cdot F_{\alpha;\, K;\, n_1+n_2-K-1}. \quad (13.3.7)$$

Правую часть выражения (13.3.7) можно записать как

$$\frac{n_1 + n_2}{n_1 \cdot n_2} \cdot T^2_{кр} = l^2.$$

В том случае, если нулевая гипотеза отвергается, то с помощью линейных контрастов можно установить причину отклонения гипотезы. Программа П 58/34 вычисляет значение статистики T^2 для каждого из линейных контрастов. Кроме этих данных программа обеспечивает вычисление верхних и нижних границ для разности $\Delta\mu$ по каждой из двух компонент. Если доверительный интервал включает в себя нулевую точку, то это означает, что по данной переменной векторы генеральных средних двух совокупностей статистически неразличимы, и наоборот.

Если необходимо построить доверительный эллипс для вектора разности $\Delta\mu$, то следует при использовании программы построения корреляционного эллипса П 52/34 вместо значения χ^2 в регистр 8 заносить значение правой части уравнения (13.3.7) l^2. Кроме того, значения выборочных стандартов s_{*x} и s_{*Y}, а также r_{*Yx} необходимо определить, используя элементы объединенной ковариационной матрицы $\mathbb{S}_*$ (13.3.3). В примере к П 58/34 мы покажем, как это сделать.

Пример. Для иллюстрации работы программы П 58/34 используем пример, приведенный в начале данного раздела.

Решение. После ввода в ПМК программы П 58/34—1 начинаем ввод данных по первой выборке: 14 ПА 225 /—/ П1 10,24 /—/ П3 0,3 /—/ П2. Далее вводим данные по второй выборке: 17 ПВ 400 /—/ П7 16 /—/ П5 0,4 /—/ П6. Вводим в ПМК компоненты вектора разности ΔP и Δf ($\Delta P = 180 - 160 = 20$; $\Delta f = 23 - 26 = -3$) : 20 П8 3 /—/ П9. Вводим служебные числа: 3 П0 4 П4. Начинаем вычисление объединенной ковариационной матрицы: В/О С/П ($-24{,}051612 = s_{*pf}$). Вычислим оценку s_{*p}, s_{*f} и r_{*pf} : ИП1 /—/ $F\sqrt{\ }$ ($17{,}915572 = s_{*p}$) ÷ ИП3 /—/ $F\sqrt{\ }$ ($3{,}6604247 = s_{*f}$) ÷ ($-3{,}6676003 \cdot 10^{-1}$) $= r_{*pf}$. Вычисляем значение статистики T^2: ИП2 С/П — результат 11,786685. Для вычисления критического значения T^2 необходимо ввести $F_{0,05;\,2;\,30} = 3{,}32$ в регистр Д : 3,32 ПД. Вычисляем $T^2_{крит}$: С/П ($6{,}861333 = T^2_{крит}$). Поскольку T^2 больше критического значения, то нулевая гипотеза о равенстве векторов отклоняется. Для вычисления данных по линейным контрастам и доверительных интервалов вводим программу П 58/34—2. Для построения корреляци-

онного эллипса вычисляем значение l^2 : В/О С/П ($8{,}3860737 \cdot 10^{-1} = l^2$). Далее С/П ($10{,}196436 = T^2_p$) С/П ($5{,}4957804 = T^2_f$). Из этих результатов видно, что причиной отклонения нулевой гипотезы послужило значимое различие векторов средних по количеству щетинок на теле моллюска. Далее С/П ($3{,}520525 \cdot 10^{-1} = \Delta f_{верх}$). С/П ($-6{,}3520525 = \Delta f_{нижн}$) С/П ($36{,}406276 = \Delta P_{верх}$) С/П ($3{,}593724 = \Delta P_{нижн}$).

Итак, линейные контрасты, а также доверительные интервалы по обеим переменным указывают на значимое различие векторов по среднему количеству щетинок. Видно, что доверительный интервал для разности длины тела моллюсков Δf включает в себя нулевое значение. Это и означает равенство сравниваемых векторов по этой переменной. А доверительный интервал для ΔP нулевого значения не включает, т. е. сравниваемые векторы значимо отличаются друг от друга именно по этой компоненте. Полученные выводы не позволяют принять нулевую гипотезу о том, что векторы морфологического критерия исследуемой популяции беспанцирных моллюсков идентичны для особей, обитающих на разных глубинах.

Используя полученные выше значения s_{*p}, s_{*f}, r_{*pf}, l^2 и программу П 52/34, вычислим данные, необходимые для построения доверительного эллипса.

Таблица 76

ΔP	36	32	28	24	20
$\Delta f_в$, мм	—3,5092608	—1,7726735	—0,8768783	—0,2753773	0,1184666
$l_в$, мм	—35,092608	—17,726735	—8,768783	—2,753773	1,184666
$\Delta f_н$, мм	—4,8886487	—6,0257578	—6,3220763	—6,3240999	—6,1184666
$l_н$, мм	—48,886487	—60,257587	—63,220763	—63,2400999	—61,184666

ΔP	16	12	8	4	3,6
$\Delta f_в$, мм	0,3240999	0,3220765	0,0257587	—1,1113513	—1,06848203
$l_в$, мм	3,240999	3,220765	0,257587	—11,113513	—10,684820
$\Delta f_н$, мм	—5,7246227	—5,1231215	—4,2273265	—2,4907392	—1,8573226
$l_в$, мм	—57,246227	—51,231215	—42,273265	—24,907392	—18,573226

Результаты вычислений приведены в табл. 76. При вводе данных по П 52/34 вместо значения $\bar{Y}$ в П3 вводим значение $\Delta f = -3$ и соответственно вместо $\bar{X}$ вводим в П5 значение $\Delta P = 20$. По адресам $53 \div 59$ вводим значение $\Delta f = 0{,}0$, а масштаб $m = 0{,}1$ заносим в регистр Д. В регистр 8 вместо χ^2 заносим $l^2 = 8{,}3860737 \cdot 10^{-1}$. Для граничного значения $\Delta P = 3{,}593724$ получаем $\Delta f_{верх} = -1{,}7700224$, а $\Delta f_{нижн} = -1{,}7711799$. Аналогично, для $\Delta P = 36{,}406276$ $\Delta f_{верх} = -4{,}2288201$, $\Delta f_{нижн} = -4{,}2299776$.

Корреляционный эллипс и доверительные интервалы для ΔP и Δf приведены на рис. 22.

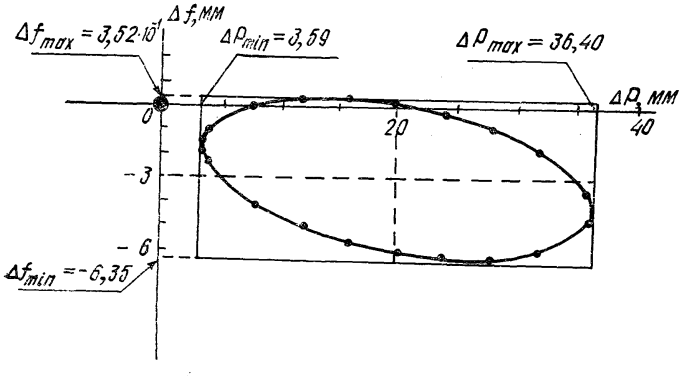

Рис. 22. Доверительный эллипс для $\Delta\mu$ и доверительные интервалы для ΔP и Δf

Отметим также некоторые особенности рассмотренных методов сравнения векторов. В обоих случаях сравнения мы имели дело с отдельными выборками, из которых получали векторы средних значений. Далее отдельные наблюдения как таковые уже не фигурировали. Однако в реальных исследованиях нередко возникают ситуации, когда имеет смысл проверить правильность априорного отнесения каждого отдельного наблюдения к той или иной группе наблюдений. При этом предполагают, что наблюдения в отдельной группе близки друг к другу по значениям переменных. Именно такую задачу разделения наблюдений и позволяет решить метод дискриминантного анализа, излагаемый в следующей главе.

ЗАДАЧИ К ГЛ. 13

1 [30, с. 78]. В табл. 77 приведены словесные и количественные оценки для выборки объемом в 10 студентов, допущенных к аспирантскому курсу по экономике. Эти студенты были выбраны из группы таких, пребы-

Таблица 77

Номер студента	1	2	3	4	5	6	7	8	9	10
Словесная оценка X_1	740	670	560	540	590	590	470	560	540	500
Количественная оценка X_2	680	600	550	520	540	700	600	540	630	600

вание которых в аспирантуре сложилось неудачно. На основании предшествующего опыта можно считать, что удовлетворительна словесная оценка, равная 590 баллам, и количественная оценка, равная 690 баллам. Следует проверить, отличаются ли средние оценки для этой группы студентов от этих значений. Величина $\alpha=5\%$.

2 [26, с. 319]. Имеется 15 значений систолического и диастолического давления крови, измеренные в мм рт. ст. Среднее значение систолического давления $\overline{X}_1=120{,}6$, диастолического — $\overline{X}_2=79{,}7$. Выборочная ковариационная матрица имеет вид

$$S=\begin{bmatrix} 438{,}26 & 343{,}02 \\ 343{,}02 & 291{,}38 \end{bmatrix}.$$

Требуется на уровне $\alpha=5\%$ провести проверку двух нулевых гипотез.

I $H_0: [\mu_1 \quad \mu_2]=[120 \quad 80]$
$H_1: [\mu_1 \quad \mu_2]\neq[120 \quad 80]$

II $H_0: [\mu_1 \quad \mu_2]=[90 \quad 60]$
$H_1: [\mu_1 \quad \mu_2]\neq[90 \quad 60]$

Принятие первой нулевой гипотезы позволит отнести эту группу к популяции здоровых лиц со средним давлением 120/80. Принятие же второй нулевой гипотезы даст основание отнести эту группу к популяции гипотоников с пониженным средним давлением 90/60. $F_{0{,}05;\,2;\,13}=3{,}8$.

3 [30, с. 89]. В табл. 78 приведены словесные и количественные оценки (в баллах) для 10 неуспевающих и 13 успевающих аспирантов, полученные ими на экзамене по экономике. На уровне $\alpha = 1\%$ требуется проверить нулевую гипотезу о равенстве двух векторов средних.

Таблица 78

Успевающие (группа 1)		Неуспевающие (группа 2)	
Словесная оценка X	Количественная оценка Y	Словесная оценка X	Количественная оценка Y
750	590	740	680
360	600	670	600
720	750	560	550
540	710	540	520
570	700	590	540
520	670	590	700
590	790	470	600
670	700	560	540
620	730	540	630
690	840	500	600
610	680	—	—
550	730	—	—
590	750	—	—
Среднее: 598,46	710,77	576,0	596,0

РЕШЕНИЯ

1. Формулируем нулевую и альтернативную гипотезы:

$$H_0 = [\mu_1 \quad \mu_2] = [590 \quad 690];$$
$$H_a : [\mu_1 \quad \mu_2] \neq [590 \quad 690].$$

Находим средние значения $\bar{X}_1 = 576$, $\bar{X}_2 = 596$, отсюда $\Delta X_1 = -14$; $\Delta X_2 = -94$.

Используя программу П 44/34, находим коэффициент корреляции $r = 3{,}9874544 \cdot 10^{-1}$; $s^2_{x_1} = 6248{,}9$ и $s^2_{x_2} = 3693{,}3$. По таблице [17, 18] находим $F_{0{,}05;\,2;\,8} = 4{,}5$.

После ввода программы П 57/34 вводим эти данные в регистры ПМК: 10 П8 3,9874544 ВП 1 /—/ П7 14 /—/ П6 94 /—/ П5 6248,9 ПВ /—/ П1 3693,3 ПА /—/ П3 4,5 П4. Вычисляем значение критерия T^2: В/О С/П ($26{,}222832 = T^2$), далее: С/П ($10{,}125 = T^2_{кр}$). Вычисляем значение $T^2_{x_2}$ для линейного контраста: С/П ($23{,}924403 = T^2_{x_2}$) и $T^2_{x_1}$: С/П ($3{,}136552 \cdot 10^{-1} = T^2_{x_1}$). Вводим компоненты вектора средних по обрабатываемой группе: 576 П5 596 П6 С/П ($657{,}15117 = \mu_{2верх}$) С/П $534{,}84883 = \mu_{2\,нижн}$) С/П ($655{,}54251 = \mu_{1\,верх}$) С/П ($496{,}45749 = \mu_{1нижн}$). Отметим, что доверительный интервал для μ_1 покрывает значение 590, тогда как доверительный интервал для μ_2 не содержит внутри себя значение 690. Это подтверждает вывод о том, что нулевая гипотеза была отвергнута из-за низких количественных оценок в данной группе. (Если принять $F_{0{,}05;\,2;\,8} = 4{,}46$, как в [30], то все результаты будут полностью идентичны приведенным в [30]).

2. После ввода программы П 57/34 вводим исходные данные. Поскольку нам явно не дано значение r, то воспользуемся для его нахождения выборочной ковариационной матрицей $\mathbb{S}$. Для этого используем выражение (13.2.13), т. е.

$$r = \frac{343{,}02}{\sqrt{438{,}26} \cdot \sqrt{291{,}38}}.$$

Несколько нарушив приведенный порядок ввода данных, вычислим r и введем его значение в Р7 в самом конце выполнения п. 2 инструкции к П 57/34. 15 П8 0,6 П6 0,3 /—/ П5 438,26 ПВ /—/ П1 291,38 ПА /—/ П3 3,8 П4. Теперь вычисляем и вводим r : 343,02 ИПВ F√ ÷ ИПА F√ ÷ ($9{,}5989483 \cdot 10^{-1} = r$) П7. Начинаем вычисление статистик: В/О С/П ($4{,}0024059 \cdot 10^{-1} = T^2$) С/П ($8{,}7846148 = T^2_{кр}$) С/П ($4{,}6331251 \cdot 10^{-3} = T^2_{x_2}$) С/П ($1{,}23221453 \cdot 10^{-2} = T^2_{x_1}$). Согласно п. 7 инструкции вводим $\bar{X}_1$ и $\bar{X}_2$: 120,6 П5 79,7 П6. Далее С/П ($92{,}309078 = \mu_{2верх}$) С/П ($67{,}090922 = \mu_{2низ}$) С/П ($136{,}06391 = \mu_{1верх}$) С/П ($105{,}13609 = \mu_{1низ}$). Поскольку $T^2 < T^2_{кр}$, то H_0 принимается. Аналогично проверяем вторую нулевую гипотезу. Результаты этой проверки следующие: $T^2 = 43{,}8789$; $T^2_{кр} = 8{,}1846148$; $T^2_{x_2} = 19{,}97855$; $T^2_{x_1} = 32{,}048099$; $\mu_{2верх} =$

$=92{,}309078$; $\mu_{2низ}=67{,}090922$; $\mu_{1верх}=136{,}06391$; $\mu_{1низ}=$ $=105{,}13609$.

Как видно из этих результатов, нулевая гипотеза о принадлежности этой группы к популяции гипотоников отклоняется.

3. Используя программу П 44/34, определим для каждой из групп студентов коэффициент корреляции между количеством баллов за словесную оценку и за количественную оценку, а также выборочные дисперсии для каждой переменной. В результате имеем для первой группы $r=3{,}0909671\cdot10^{-1}$; $s^2_X=10130{,}769$; $s^2_Y=4707{,}6923$; для второй группы $r=3{,}9873476\cdot10^{-1}$; $s^2_X=6248{,}8888$; $s^2_Y=3693{,}3333$.

Далее используем программу П 58/34—1, после ввода которой вводим исходные данные. Вводим данные по первой группе: 12 ПА 10130,769 /—/ П1 4707,6923 /—/ П3 3,0909671 ВП 1 /—/ П2. Аналогично вводим данные по второй группе: 9 ПВ 6248,8888 /—/ П7 3693,3333 /—/ П5 3,9873476 ВП 1 /—/ П6. Вводим вектор разности средних: $\Delta\bar{X}=598{,}46-576=22{,}46$ и $\Delta Y=710{,}77-596=$ $=114{,}77$: 22,46 П8 114,77 П9. Вводим вспомогательные числа: 3 П0 4 П4. Начинаем вычисление элементов объединенной ковариационной матрицы S_* : В/О С/П ($2040{,}7324=s_{*XY}$). Выполняем п. 5 инструкции: ИП1 /—/ $\sqrt{F}$ ($92{,}016877=s_{*X}$) ÷ ИП3 /—/ $\sqrt{F}$ ($65{,}367935=$ $=s_{*Y}$) ÷ ($3{,}3927651\cdot10^{-1}=r_{*XY}$). Вычисляем значения статистики T^2—Хотеллинга: ИП2 С/П ($18{,}213424=T^2$). Далее вводим табличное значение $F_{0{,}01;\,2;\,20}=5{,}85$ в РД : 5,85 ПД и вычисляем значение критическое статистики T^2 : С/П ($12{,}285=T^2_{крит}$).

Поскольку $T^2>T^2_{крит}$, то нулевая гипотеза о равенстве двух векторов средних отвергается. Исходя из этого представляет интерес использование линейных контрастов для того, чтобы определить причину отклонения нулевой гипотезы. Не выключая ПМК, вводим программу П 58/34—2 и выполняем п. 10 инструкции: В/О С/П ($2{,}1735=l^2$). Далее С/П ($17{,}423793=T^2_Y$) С/П ($3{,}3674413\cdot$ $\cdot10^{-1}=T^2_X$). Поскольку $T^2_Y>T^2_{крит}$ и $T^2_X<T^2_{крит}$, то делаем однозначный вывод о том, что на заданном уровне значимости $\alpha=1\%$ словесные оценки этих двух групп статистики не различимы, тогда как количественные

оценки этих двух групп достоверно различаются. Именно это различие и является причиной отклонения нулевой гипотезы. Далее вычисляем границы доверительных интервалов для $\Delta\bar{Y}$ и $\Delta\bar{X}$: С/П (211,14061 = $\Delta Y_{верх}$) С/П (18,3994 = $\Delta Y_{нижн}$) С/П (158,11859 = $\Delta X_{верх}$) С/П (—113,19859 = $\Delta X_{нижн}$).

Доверительные интервалы также подтверждают сделанный ранее вывод: для ΔY доверительный интервал не содержит внутри себя нулевое значение, а для ΔX— содержит.

ГЛАВА 14

ИССЛЕДОВАНИЕ НАБЛЮДЕНИЙ МЕТОДОМ ДИСКРИМИНАНТНОГО АНАЛИЗА

Интеллект состоит в том, чтобы узнать подобие разных вещей и разницу подобных.

Ш. Монтескье

14.1. Понятие о дискриминантном анализе

Основной задачей дискриминантного анализа является разделение, дискриминация исследуемого множества объектов, характеризующихся набором закоррелированных выходных характеристик, на две или большее число групп. В результате такой дискриминации в одной группе оказываются наблюдения (объекты), наиболее близкие по своим свойствам, тогда как межгрупповое отличие максимизируется. Для иллюстрации этого положения обратимся к рис. 23.

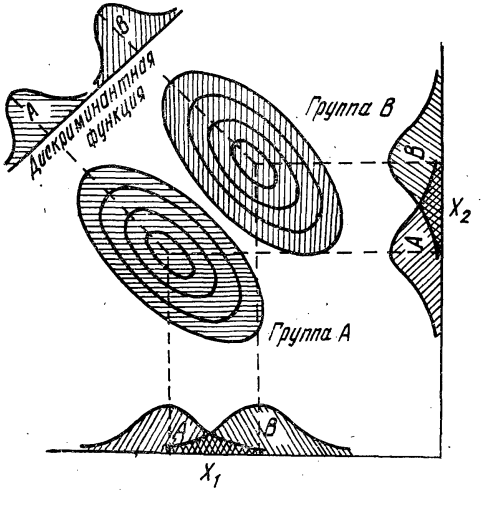

Рис. 23. Разделение объектов на две группы с помощью дискриминантной функции

Для удобства визуального восприятия на этом рисунке изображены двухмерные наблюдения, т. е. имеющие две закоррелированных характеристики X_1 и X_2. При проекции этих двух групп А и В на оси X_1 и X_2 происходит их перекрытие, наложение. Для того чтобы надежно отнести каждый объект к группе А или В, необходимо определить такое направление, проекция на которое давало бы максимальное разделение этих двух групп. На рисунке изображена линия дискриминантной функции, которая и позволяет решать подобную задачу дискриминации исследуемых объектов. Естественно, что сложность подобного разделения многократно возрастает по мере повышения размерности вектора характеристик свойств исследуемых объектов.

Дискриминантный анализ был разработан независимо друг от друга известными статистиками Р. Фишером, Г. Хотеллингом и Махаланобисом.

В зависимости от вида дискриминантных функций различают линейный дискриминантный анализ и нелинейный дскриминантный анализ. Не рассматривая подробно процедуру вывода выражения для коэффициентов линейной дискриминантной функции, мы приведем конечные выражения и обсудим их. Затем будет приведена программа линейного дискриминантного анализа двухмерных объектов исследования.

14.2. Линейный дискриминантный анализ двухмерных наблюдений

Выражения, которые далее используются, мы уже применяли при проверке гипотез о векторах средних. Так, S_* — это объединенная ковариационная матрица, определяемая по (13.3.3), S_*^{-1} — матрица, обратная объединенной ковариационной матрице $S_* \cdot [\overline{X}_1 - \overline{X}_2]$ — матрица разности векторов средних первой и второй выборки группы. Поскольку в реальных ситуациях мы имеем дело с ограниченными выборками из генеральных совокупностей, то получаемые результаты являются оценкой дискриминантной функции. Цель проводимого анализа заключается в получении оценок коэффициентов линейной функции

$$\varphi_{il} = \lambda_1 X_{i1l} + \lambda_2 X_{i2l}, \quad i = 1, 2, \quad l = 1, 2, \ldots, n_i, \qquad (14.2.1)$$

которая и является линейной дискриминантной функцией. Здесь индекс i относится к группе, а индекс l — номер наблюдения в группе. Оценкой неизвестных коэффициентов функции (14.2.1) является матрица

$$\hat{\lambda} = S^{-1}_{*} \cdot (\bar{X}_1 - \bar{X}_2) . \qquad (14.2.2)$$

Видно, что правая часть выражения (14.2.2) является частью выражения (13.3.5). Произведение трех матриц в (13.3.5) называется расстоянием Махаланобиса D^2 или обобщенным расстоянием:

$$D^2 = (\bar{X}_1 - \bar{X}_2) \cdot S^{-1}_{*} \cdot (\bar{X}_1 - \bar{X}_2). \qquad (14.2.3)$$

Это мера разделения между двумя многомерными средними, выраженная в единицах дисперсии объединенной выборки. Критерий Хотеллинга T^2 выражается через расстояние D^2 следующим образом:

$$T^2 = \frac{n_1 \cdot n_2}{n_1 + n_2} \cdot D^2 . \qquad (14.2.4)$$

Выражение (14.2.4) используется для проверки значимости дискриминантной функции. Этот критерий проверяется так же, как и при проверке гипотезы о равенстве векторов средних. Проверяемая нулевая гипотеза в данном случае заключается в том, что два многомерных средних равны между собой или что расстояние между ними равно нулю. В чем же отличие этих двух методов анализа? При анализе векторов средних мы проверяем гипотезу относительно векторов средних, не рассматривая далее каждое наблюдение в отдельности. В дискриминантном же анализе мы, получив оценку дискриминантной функции и подставляя в нее значения X_{i1l} и X_{i2l}, можем заново, более точно произвести переклассификацию наблюдений, относя их в зависимости от значения функции к той или другой группе. В этом и заключается основное отличие — в возможности апостериорной переклассификации наблюдений.

Относительный вклад j-той переменной в расстояние между средними двух групп измеряется величиной [33]:

$$E_j = \frac{\lambda_j D_j}{D^2} , \qquad (14.2.5)$$

где D_j — разность между средними по j-той переменной. Эта мера является полным аналогом стандартизованных коэффициентов регрессии в уравнениях множественной регрессии.

Каким же образом, зная оценки коэффициентов λ_i, можно производить переклассификацию наблюдений? В том случае, когда цена ошибочной классификации одинакова для обеих групп и если априорные вероятности принадлежности наблюдения к каждой из этих групп одинаковы, то выбирается граничное значение функции φ_0, которое будет находиться посредине между φ_1 и φ_2, т. е.

$$\varphi_0 = \frac{1}{2} \cdot \hat{\lambda} \cdot [\bar{X}_1 + \bar{X}_2] . \qquad (14.2.6)$$

Правило классификации в этом случае можно сформулировать в следующей форме: наблюдение отнести к группе 1, если $\varphi_{i1} \geqslant \varphi_0$; наблюдение отнести к группе 2, если $\varphi_{i1} < \varphi_0$. Используя выражение (14.2.6), правило классификации можно сформулировать иначе: наблюдение отнести к группе 1, если

$$\varphi = \hat{\lambda} X^T - \frac{1}{2} \hat{\lambda}^T (\bar{X}_1 + \bar{X}_2) \geqslant 0,$$

где $$X^T = [\, X^T_1 \,\vdots\, X^T_2 \,] ; \qquad (14.2.7)$$

наблюдение отнести к группе 2, если

$$\varphi = \hat{\lambda} X^T - \frac{1}{2} \lambda^T (\bar{X}_1 + \bar{X}_2) < 0 . \qquad (14.2.8)$$

Ниже приведена программа П 59/34, которая позволяет произвести оценку коэффициентов дискриминантной функции для случая двух выборок и двух переменных. Исходные данные для работы этой программы полностью идентичны данным, используемым при сравнении двух векторов средних. Производится также вычисление статистики T^2, критического значения этой статистики, расстояния Махаланобиса D^2, значения φ_0.

Пример. Для иллюстрации работы программы П 59/34 используем данные предыдущего примера к программе П 58/34, когда мы проверяли гипотезу о равенстве двух векторов средних. Отклонение нулевой гипотезы о равенстве этих векторов побудило провести более деталь-

ное исследование моллюсков из обеих выборок. Возможно, что в каждой из выборок имеются особи из большей и меньшей глубин обитания. Попытаемся с помощью дискриминантного анализа произвести переклассификацию моллюсков уже не по признаку «глубина срабатывания трала», а по их морфологическим признакам P и f.

Решение. Набрав программу П 59/34 и осуществив ввод исходных данных задачи согласно инструкции к П 59/34, вычисляем оценки коэффициентов дискриминантной функции: В/О С/П ($5{,}2610285 \cdot 10^{-2} = \lambda_p$) С/П ($-1{,}2946306 \cdot 10^{-1} = \lambda_f$) С/П ($1{,}4405949 = D^2$). Для вычисления значения φ_0 вводим в ПМК два вектора средних: 180 П4 160 П5 23 П6 26 П7. Далее С/П ($5{,}7719035 = \varphi_0$). Итак, левая часть выражений (14.2.7) и (14.2.8) для переклассификации наблюдений имеет следующий вид:

$$\varphi = -5{,}7719035 + 5{,}2610285 \cdot 10^{-2} \cdot P_{i1} - 0{,}12946306 \cdot f_{i1}. \qquad (14.2.9)$$

Оставшаяся часть программы и производит вычисление значения φ по (14.3.9). Однако прежде чем производить переклассификацию, необходимо проверить достоверность, значимость полученной оценки дискриминантной функции по T^2-критерию. Для этого вначале вычислим значение T^2, нажимая клавиши ПМК в следующей последовательности: ИПА 1+П4 ИПВ 1+П5× ИП4 ИП5+÷ИП1× ($11{,}786685 = T^2$). Далее вычисляем критическую величину T^2, имея в виду, что $F_{0{,}05;\,2;\,30} = 3{,}32$. Нажимаем клавиши в следующей последовательности: ИПС ИПС 1— ÷2 × 3,32× ($6{,}8613331 = T^2_{кр}$). Поскольку T^2 больше критического значения $T^2_{кр}$, то полученную оценку дискриминантной функции считаем достоверной.

Покажем далее на нескольких особях процедуру переклассификации:

Моллюск № 1 : P = 180, f = 25 мм
Моллюск № 2 : P = 175, f = 28 мм
Моллюск № 3 : P = 170, f = 26 мм

Выполняем: 180↑25 С/П ($4{,}613713 \cdot 10^{-1} > 0$)
175↑28 С/П ($-1{,}900693 \cdot 10^{-1} < 0$)
170↑26 С/П ($-1{,}951946 \cdot 10^{-1} < 0$)

Таким образом, в зависимости от знака получаемого при этом значения φ мы классифицируем отдельные особи к группе 1 или к группе 2.

Какова цель, преследуемая при использовании дискриминантного анализа? Очевидно, что главной целью этого вида анализа является классификация объектов исследования на однородные группы. Исследуя неоднородные выборки, мы не получим достаточно точные и надежные результаты по объективно существующим зависимостям между характеристиками этих объектов для каждой из этих групп. Проведя разделение выборок на однородные группы, мы не только повысим точность и надежность результатов исследования, но и сможем установить причины возникновения неоднородности выборок.

ЗАДАЧИ К ГЛ. 14

1 [30, с. 226]. Провести дискриминантный анализ данных задачи 3 (глава 13). Уровень значимости принять равным 1%. После нахождения дискриминантной функции произвести переклассификацию по обеим группам.

2. Исследовались механические свойства пряжи. Для этого в двух партиях пряжи ($n_1=11$, $n_2=12$) измерялись крепость пряжи (в граммах), определяемая величиной предельной нагрузки (X), при которой пряжа не разрывается, и удлинение пряжи (в %) — отношение приращения длины при предельной нагрузке к первоначальной длине образца (Y). В результате получены следующие значения статистик:

$\bar{X}_1=300$; $s^2_{1(x)}=400$; $\bar{Y}_1=10$; $s^2_{1(Y)}=1{,}2$; $r_1=0{,}5$;

$\bar{X}_2=320$; $s^2_{2(x)}=900$; $\bar{Y}_2=12{,}1$; $s^2_{2(Y)}=4$; $r_2=0{,}69$.

Требуется определить вид дискриминантной функции и проверить ее достоверность при $\alpha=5\%$.

3. Изменяя в условиях задачи 2 значение $s^2_{2(x)}$ от 900 до 1600, установите, при каком значении $s^2_{2(x)}$ дискриминантная функция перестанет быть достоверной при $\alpha=5\%$. Попытайтесь дать собственную интерпретацию этой ситуации.

РЕШЕНИЯ

1. После ввода программы П 59/34 выполняем п. 2 и 3 инструкции: 12 ПА 10130,769 /—/ П1 4707,6923 /—/

П3 3,0909671 ВП 1 /—/ П2 9 ПВ 6248,8888 /—/ П7 3693,3333 /—/ П5 3,9873476 ВП 1 /—/ П6 22,46 П8 114,77 П9 3 П0 4 П4. Вычисляем оценки коэффициентов дискриминантной функции λ_x и λ_Y : В/О С/П ($-4{,}3180918\cdot10^{-3}=\lambda_x$) С/П ($2{,}892184\cdot10^{-2}=\lambda_Y$) С/П ($3{,}2223752=D^2$). Вычислим значение φ_0 : 598,46 П4 576 П5 710,77 П6 596 П7 С/П ($16{,}361383=\varphi_0$).

Далее вычислим значение статистики T^2 : ИПА 1 +П4 ИПВ 1+П5×ИП4 ИП5×÷ИП1× ($18{,}213425=T^2$). Далее ИПС ИПС 1 — ÷2×5,85× ($12{,}285=T^2_{крит}$). Как видим, T^2 и $T^2_{крит}$ при $F_{0{,}01;\,2;\,20}=5{,}85$ равны значениям T^2 и $T^2_{крит}$, полученным при решении задачи 3 (глава 13).

Приступаем к переклассификации наблюдений по обеим группам. Вначале по группе 1 : 750↑590 С/П ($-2{,}536066=\varphi_{11}<0$) 360↑600 С/П ($-5{,}62792\cdot10^{-1}=\varphi_{12}<0$) 720↑750 С/П ($2{,}220971=\varphi_{13}>0$). Значения φ_{1i} для остальных наблюдений 1-й группы также положительны. Значит первые два наблюдения этой группы ошибочно отнесены в данную группу. На основании переклассификации их следует отнести во вторую группу. Аналогично во второй группе для всех наблюдений, кроме 1-го ($\varphi_{21}=1{,}1008\cdot10^{-1}>0$) и 6-го ($\varphi_{26}=1{,}336231>0$), будут отрицательны, т. е. первое и шестое наблюдение второй группы следует отнести к первой группе.

2. Выполнив полностью инструкцию к программе П 59/34, получим следующие значения требуемых статистик (идентифицируйте их самостоятельно): 1,6573486; $-8{,}181239\cdot10^{-1}$; $3{,}035581\cdot10^{-3}$; $-8{,}099239$; 9,5117397. При значении $F_{0{,}05;\,2;\,20}=3{,}5$ $T^2_{кр}=7{,}35$.

Выполнив идентификацию статистик, запишите окончательный вид дискриминантной функции $\varphi=f(X, Y)$. Может ли быть использована эта функция для дискриминации отдельных наблюдений при заданном значении уровня значимости α?

ГЛАВА 15

ПРИМЕНЕНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ (МГК) ДЛЯ АНАЛИЗА НАБЛЮДЕНИЙ

Идите вперед и уверенность придет.

Д'Аламбер

15.1. Основные особенности МГК

Желание исследователя повысить точность и глубину познания исследуемого явления или процесса неизбежно приводит его к увеличению общего числа характеристик, используемых для их описания. Такая многомерность имеет не только положительные стороны, но и отрицательные. Во все возрастающем числе связей и зависимостей все труднее выделить основные, доминирующие закономерности. Возникает потребность в «сжатии», редукции имеющейся информации. Одним из методов такой редукции является понижение размерности исходного признакового пространства путем перехода к новым, обобщенным переменным, причем независимым между собой. Такими новыми переменными и являются главные компоненты. Метод главных компонент называют также компонентным анализом. Большой вклад в развитие компонентного анализа внесли такие ученые, как Хотеллинг, Спирмэн, Харман, Фишер, Махаланобис и Рао. Какие же задачи решаются с помощью этого метода многомерного статистического анализа [3, 4, 26, 30—37].

1. Обнаружение скрытых, но объективно существующих закономерностей, которые обусловлены воздействием как внутренних, так и внешних причин.

2. Описание исследуемого процесса числом главных компонент m, значительно меньшим, чем число исходных признаков. Этим достигается редукция имеющейся информации с минимальными потерями. Следует отметить, что выделенные главные компоненты содержат в среднем больше информации, чем отдельные непосредственно измеряемые признаки.

3. Выявление и изучение стохастических связей признаков с главными компонентами, что позволяет определить исходные признаки, наиболее тесно связанные с найденными главными компонентами.

4. Использование полученных результатов для прогнозирования развития процесса, явления на основе знания величины отдельных главных компонент.

Для упрощения выкладок рассмотрим трехмерный случай.

При измерении трех нормально распределенных параметров X, Y и Z, которые определенным образом закоррелированы между собой, n наблюдений (n точек) разместятся в трехмерном пространстве с тремя осями X, Y и Z в виде облака вокруг общего центра тяжести. Это облако в общем случае имеет форму эллипсоида. В частном случае, когда во всех трех направлениях дисперсия одинакова по величине, эллипсоид вырождается в шар.

Известно, что эллипсоид имеет три главные оси, которые пересекаются в центре тяжести. Самый длинный диаметр овального тела является первой главной осью

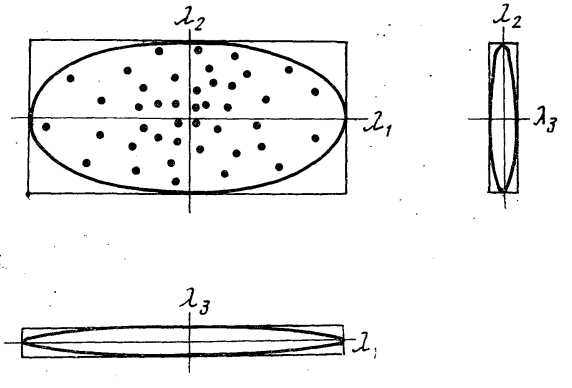

Рис. 24. Главные оси эллипсоида

λ_1. Второй главной осью λ_2 является самый длинный диаметр в плоскости, ортогональной к первой главной оси и проходящей через центр тяжести. Третья главная ось λ_3 в трехмерном эллипсоиде перпендикулярна первой и второй главным осям и также проходит через центр (рис. 24).

Основная идея метода главных компонент состоит в переходе к новой системе координат, которая совпадает с главными осями эллипсоида, и оценке важности, информативности каждой из новых переменных — главных компонент. Компоненты, несущие малую информативную нагрузку, далее отбрасываются и не рассматриваются. Ниже приводится последовательность использования метода главных компонент (МГК) и его программная реализация для 3-мерного случая.

15.2. Оценка главных компонент

Очевидно, что новые переменные, совпадающие с главными осями эллипсоида, являются линейными комбинациями старых переменных X, Y и Z. Если обозначить, например, первую главную компоненту λ_1, то для i-го наблюдения можно записать

$$\lambda_{1i}=\alpha_1 X_i+\beta_1 Y_i+\gamma_1 Z_i, \qquad (15.2.1)$$

где X_i, Y_i и Z_i — координаты i-той точки на старые оси X, Y и Z. Когда эллипсоид сильно вытянут по одним осям и сдавлен по другим, то появляется возможность пренебречь одной из осей с незначительной потерей информации. Так, на рис. 24 $\lambda_1>\lambda_2\gg\lambda_3$, поэтому с небольшой потерей информации от этих трех главных компонент можно перейти к двум: λ_1 и λ_2. В этом случае вместо трехмерного эллипсоида мы рассматриваем только его проекцию на плоскость λ_1—λ_2, т. е. двухмерный эллипс. Отметим, что главные компоненты ортогональны, т. е. они не коррелированы между собой ($\rho_{12}=0$).

Теперь рассмотрим вопрос о размерностях переменных X, Y и Z в выражении (15.2.1). Очевидно, что непосредственная подстановка координат каждого наблюдения в (15.2.1) возможна только тогда, когда размерности этих переменных идентичны. В противном случае, например, метры будут складываться с килограммами, рублями и т. п., что затрудняет содержательную интерпретацию. Поэтому когда переменные X, Y и Z имеют различные единицы измерения, то предварительно коор-

динаты наблюдений необходимо стандартизовать по известным формулам:

$$\tilde{X}_i = \frac{X_i - \bar{X}}{s_X}; \quad \tilde{Y}_i = \frac{Y_i - \bar{Y}}{s_Y};$$

$$\tilde{Z}_i = \frac{Z_i - \bar{Z}}{s_Z}. \qquad (15.2.2)$$

Стандартизованные переменные являются безразмерными переменными и их уже можно складывать между собой.

Форма эллипсоида для стандартизованных переменных (его вытянутость и сплющенность по тем или иным главным осям, а также расположение этих осей относительно старой системы координат) полностью определяется набором коэффициентов корреляции ρ_{YX}, ρ_{XZ} и ρ_{YZ}. Ниже мы везде опускаем знак «~», подразумевая, что мы имеем дело со стандартизованными переменными.

Первоначально по выборке трехмерных наблюдений определяются коэффициенты корреляции r_{YX}, r_{XZ} и r_{YZ}. Далее для нахождения коэффициентов α, β и γ в (15.2.1) используется метод собственных значений и собственных векторов. Для нахождения собственных чисел корреляционной матрицы необходимо решить характеристическое уравнение

$$\Delta = \begin{vmatrix} 1-\lambda & r_{XY} & r_{XZ} \\ r_{YX} & 1-\lambda & r_{YZ} \\ r_{XZ} & r_{ZY} & 1-\lambda \end{vmatrix} = 0, \qquad (15.2.3)$$

которое можно записать иначе:

$$\lambda^3 - I\lambda^2 + J\lambda - D = 0. \qquad (15.2.4)$$

Здесь I, J и D так называемые инварианты относительно параллельного переноса и поворота осей X, Y и Z. Уравнение третьей степени имеет три корня, которые связаны с I, J и D: $I=\lambda_1+\lambda_2+\lambda_3$; $J=\lambda_1\cdot\lambda_2+\lambda_2\cdot\lambda_3+\lambda_3\cdot\lambda_1$; $D=\lambda_1\cdot\lambda_2\cdot\lambda_3$. Для нахождения корней уравнения третьей степени первоначально с использованием корреляционной матрицы находим коэффициенты I, J, D.

Для нашего случая представляет интерес только тот случай, когда все три корня λ_1, λ_2, λ_3 положительны. Только в этом случае квадратичная форма от трех ар-

гументов положительно определена и описывает эллипсоид. Если же один из трех корней есть отрицательное число, то в этом случае мы имеем дело с гиперболической квадратичной формой.

Каждому из трех собственных значений λ_i соответствует свой собственный вектор, определяемый выражением, аналогичным (15.2.1). Обычно в качестве собственных векторов принято использовать ортонормированные векторы, для которых выполняются следующие равенства:

$$\alpha_i\alpha_j + \beta_i\beta_j + \gamma_i\gamma_j = 0, \ i \neq j; \qquad (15.2.5)$$

$$\alpha_i\alpha_j + \beta_i\beta_j + \gamma_i\gamma_j = 1, \ i = j. \qquad (15.2.6)$$

Условие нормировки (15.2.6) можно записать иначе:

$$\alpha_i^2 + \beta_i^2 + \gamma_i^2 = 1.$$

Условие (15.2.5) отражает ортогональность собственных векторов.

Остановимся на некоторых свойствах характеристических корней λ_i.

1. Сумма характеристических корней равна сумме элементов главной диагонали корреляционной матрицы R. Поскольку на главной диагонали матрицы R находятся $r_{ii} \equiv 1$, то для случая трех переменных $\lambda_1 + \lambda_2 + \lambda_3 = I = 3$.

2. Произведение характеристических корней равно определителю D корреляционной матрицы R, т. е. $\lambda_1 \cdot \lambda_2 \cdot \lambda_3 = D = |R|$. Информативность, важность i-той главной компоненты можно оценить следующим выражением:

$$K_i(\%) = \frac{100 \cdot \lambda_i}{3} \%, \qquad (15.2.7)$$

где K_i — информативность i-той главной компоненты.

Реализация МГК на ПМК 34 занимает две последовательные программы. Исходными данными для первой программы являются выборочные коэффициенты парной корреляции r_{xY}, r_{xz} и r_{Yz}. В результате работы первой программы вычисляются коэффициенты уравнения (15.2.3), решается уравнение (15.2.4) и для всех трех корней λ_1, λ_2 и λ_3 вычисляются информативности K_i по (15.2.7). Кроме того, вычисляется суммарная информативность первых двух главных компонент. С помощью

второй программы вычисляются значения α_i, β_i и γ_i для первой и второй главных компонент. Далее с помощью значений α, β и γ производится вычисление значений первой и второй главных компонент для отдельных наблюдений.

Поскольку исходными данными для МГК являются выборочные коэффициенты корреляции, то, очевидно, необходимо проверить значимость, достоверность полученных результатов. Рассмотрим особенности формулируемой при этом статистической гипотезы.

Как уже говорилось выше, оценка информативности отдельных главных компонент связана с величиной отдельных характеристических корней — корней уравнения (15.2.4). Поэтому основная проверяемая гипотеза формулируется как гипотеза проверки различия отдельных корней между собой. Предположим, что для нашего трехмерного случая мы определили характеристические корни $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3$. В случае равенства всех трех корней эллипсоид превращается в сферу. В двухмерном же случае эллипс превращается в окружность. В этом случае нет никакого преимущества любых новых осей по сравнению с исходными осями X, Y и Z, т. к. информативность новых осей не увеличивалась. Следовательно, применение МГК не имеет смысла, т. к. невозможно уменьшить число переменных.

Поэтому нулевая гипотеза H_0 при использовании МГК формулируется как гипотеза проверки сферичности исходного массива наблюдений:

$$H_0 : \lambda_1 = \lambda_2 = \lambda_3, \qquad (15.2.8)$$

$$H_1 : \text{не все } \lambda_i \text{ равны между собой.}$$

Установлено, что для проверки нулевой гипотезы H_0 можно использовать следующую статистику [30]:

$$\chi^2 = -(n-1) \cdot \ln|R|, \qquad (15.2.9)$$

где $|R|$ — определитель корреляционной матрицы R. Напомним, что $|R| = D$ из (15.2.4). Статистика χ^2 по (15.2.9) имеет χ^2-распределение с числом степеней свободы $K = \frac{m(m-1)}{2}$, где m — число исходных переменных. Для нашего случая с тремя переменными $K = 3$.

Таким образом, используя (15.2.9), мы можем оценить значимость всей корреляционной матрицы в целом, а не по отдельным r_{ij}. Поскольку $|R|=D$, то уже в самом начале работы первой программы П 60/34 мы можем проверить целесообразность дальнейших вычислений и применения МГК.

15.3. Применение МГК в регрессионном анализе

МГК используется не только для уменьшения числа переменных. Весьма интенсивно развивается в последние годы использование главных компонент в регрессионных уравнениях при наличии мультиколлинеарности независимых переменных X_i. Рассмотрим идею метода.

Для переменных X_i вычисляют главные компоненты и ограничиваются некоторым числом компонент. Далее для зависимой переменной Y находится уравнение регрессии от оставленных главных компонент. После этого переходят к уравнению регрессии Y от исходных переменных X_i [31]. Предположим, что имеется четыре переменных: Y, X_1, X_2, X_3. При этом переменные X_1, X_2 и X_3 связаны мультиколлинеарностью. Поэтому, как уже было рассмотрено в главе 11, уравнение регрессии $Y=f(X_1, X_2, X_3)$ будет весьма неустойчиво к изменениям переменных X_1, X_2 и X_3. Вычисляем главные компоненты λ_1, λ_2 и λ_3. Предположим, что первые две главные компоненты вместе имеют достаточно высокую информативность и поэтому ограничиваемся этими двумя компонентами. Напомним, что главные компоненты ортогональны, т. е. мы от закоррелированных переменных X_1, X_2 и X_3 перешли к независимым, незакоррелированным переменным λ_1 и λ_2:

$$\lambda_1 = \alpha_1 X_1 + \beta_1 X_2 + \gamma_1 X_3 , \qquad (15.3.1)$$
$$\lambda_2 = \alpha_2 X_1 + \beta_2 X_2 + \gamma_2 X_3 .$$

В этом случае уравнение регрессии Y на главные компоненты λ_1 и λ_2 имеет следующий вид:

$$Y = b_0 + b_1\lambda_1 + b_2\lambda_2 = b_0 + b_1(\alpha_1 X_1 + \beta_1 X_2 + \gamma_1 X_3) + b_2(\alpha_2 X_1 + \beta_2 X_2 + \gamma_2 X_3) + e = b_0 + (b_1\alpha_1 + b_2\alpha_2)X_1 + (b_1\beta_1 + b_2\beta_2)X_2 + (b_1\gamma_1 + b_2\gamma_2) \cdot X_3 . \qquad (15.3.2)$$

В том случае, если бы мы оставили все три главные компоненты, то коэффициенты при переменных X_i полностью совпали бы с коэффициентами, найденными прямой регрессией Y на все переменные X_i.

Поскольку главные компоненты некоррелированы, то это облегчает процедуру пошагового отбора информативных для предсказания главных компонент [3, с. 254].

Пример [57, с. 688]. Проводилось экспериментальное исследование раствора, содержащего два склонных к осмолению вещества. По результатам 35 измерений установлены коэффициенты корреляции: $r_{Yx_1}=0{,}3673$, $r_{Yx_2}=0{,}5409$ и $r_{x_1x_2}=0{,}2504$, где Y — вязкость раствора, сП; X_1 — концентрация вещества A_1, г/л; X_2 — концентрация вещества A_2, г/л. Исследовать имеющиеся данные с помощью МГК.

Решение. Вводим программу П 60/34—1 и вспомогательные числа по п. 2 инструкции: 3 /—/ П2 4 ПА 0,03 ПД. Вводим коэффициенты корреляции: 0,3673 П7 0,5409 П8 0,2504 П9. Вычисляем коэффициенты уравнения (15.2.4): В/О С/П (2,5098177=J). Извлекаем из регистров 0 и 2 значения остальных коэффициентов: ИПО ($-6{,}09313\cdot10^{-1}=-$D), ИП2 (—3=—I). Таким образом, уравнение (15.2.4) в данном случае имеет следующий вид:

$$\lambda^3-3\lambda^2+2{,}5098177\lambda-6{,}09313\cdot10^{-1}=0.$$

Прежде чем проводить дальнейшие вычисления, проведем проверку нулевой гипотезы о сферичности исходного массива наблюдений, т. е. гипотезы $H_0:\lambda_1=\lambda_2=\lambda_3$. Для этого вручную выполним следующие команды для вычисления χ^2 по (15.2.9): ИПО /—/ Fln /—/ 34$\times$(16,844389=χ^2). По таблице [17, 18] находим $\chi^2_{0{,}05;\,3}=7{,}82$. Поскольку $\chi^2>\chi^2_{0{,}05;\,3}=7{,}82$, то нулевую гипотезу о равенстве характеристических корней уравнения (15.2.4) отклоняем. Продолжаем вычисления по п. 5 инструкции: С/П (85,28803% $=K_1+K_2$). Извлекаем из регистров памяти значения λ_i и K_i: ИПС (1,7854059$=\lambda_1$) ИПВ ($7{,}7323502\cdot10^{-1}=\lambda_2$) ИПА ($4{,}4135908\cdot10^{-1}=\lambda_3$) ИП4 (59,51353% $=K_1$) ИП5 (25,7745% $=K_2$) ИП6 (14,711969$=K_3$) ИП3 (85,28803$=K_1+K_2$).

Таким образом, две главные компоненты несут примерно 85% всей исходной информации о растворе, содержащейся в 35 измерениях трех переменных: X_1, X_2 и Y.

Вводим вторую программу П 60/34—2. После этого выполняем п. 7 и 8 инструкции: 0 П6 В/О С/П ($4{,}9853295 \cdot 10^{-1} = \gamma_2$). Извлекаем из регистров значения коэффициентов:

ИП 1 ($6{,}3824049 \cdot 10^{-1} = \alpha_1$) ИП 2 ($4{,}8823886 \cdot 10^{-1} = \beta_1$)
ИП3 ($5{,}9520746 \cdot 10^{-1} = \gamma_1$) ИП4 ($1{,}8323369 \cdot 10^{-1} = \alpha_2$)
ИП5 ($-8{,}4728405 \cdot 10^{-1} = \beta_2$) ИП6 ($4{,}9853295 \cdot 10^{-1} = \gamma_2$).

Таким образом, значения первой и второй главных компонент можно вычислять по формулам:

$$\lambda_{1i} = 0{,}638 \cdot Y_i + 0{,}488 \cdot X_{1i} + 0{,}595 \cdot X_{2i};$$
$$\lambda_{2i} = 0{,}183 \cdot Y_i - 0{,}847 \cdot X_{1i} + 0{,}499 \cdot X_{2i}.$$

ЗАДАЧИ К ГЛАВЕ 15

1 [37, с. 155]. По данным измерений (мм) длины (X_1), ширины (X_2) и высоты (X_3) панциря 24 особей одного из видов черепах определены коэффициенты корреляции: $r_{12} = 0{,}9740$; $r_{13} = 0{,}9726$; $r_{23} = 0{,}9655$.

Используя МГК, определить новые оси λ_1 и λ_2.

2 [23, с. 223]. При исследовании экономических показателей 12 предприятий получены коэффициенты корреляции $r_{12} = 0{,}2678$; $r_{13} = 0{,}7426$; $r_{23} = 0{,}7245$, где X_1 — объем капиталовложений; X_2 — средний процент выполнения нормы; X_3 — объем производства. Используя МГК, определить новые оси компонент λ_1, λ_2, λ_3.

3 [45, с. 129, 135]. При исследовании взаимосвязи максимального артериального давления крови человека (X_1), его возраста (X_2) и веса (X_3) были определены коэффициенты корреляции: $r_{12} = -0{,}59$; $r_{13} = -0{,}41$; $r_{23} = 0{,}91$. Используя программу П 60/34—1, определить значения λ_i и K_i. Проверить гипотезу сферичности данных при $n = 25$.

4 [55, с. 153]. По двум участкам (№ 1 и Малотаскинский) Теченского железорудного месторождения были проанализированы 342 керновых пробы на содержание различных элементов: серы (X), железа (Y) и кобальта (Z). На основе этих анализов получены следующие коэффициенты корреляции: $r_{xY} = 0{,}420$; $r_{xz} = 0{,}684$; $r_{Yz} = 0{,}429$. Использовать МГК и проверить гипотезу сферичности данных при $\alpha = 5\%$.

5 [56, с. 103]. При изучении 13 скважин, пробуренных на одном из уральских месторождений меди, получены данные о содержании меди (X), серы (Y) и цинка (Z): $r_{xY}=0{,}47$; $r_{xz}=-0{,}15$; $r_{Yz}=-0{,}27$. Используя МГК, проверьте гипотезу сферичности этих данных при $\alpha=5\%$.

6. Используя значения коэффициентов корреляции предыдущих задач, попытайтесь составить условия задач по собственной тематике исследований.

7. Изменяя величину одного из коэффициентов корреляции в задаче 1, проанализируйте поведение величин λ_i и K_i. При некоторых значениях r_{ij} задача не будет иметь решения. Дайте содержательную интерпретацию таких случаев.

РЕШЕНИЯ

1. После ввода программы П 60/34—1, ввода вспомогательных чисел и коэффициентов корреляции вычисляем коэффициенты уравнения (15.2.4), которое имеет вид

$$\lambda^3-3\lambda^2+1{,}73183\cdot10^{-1}\cdot\lambda-2{,}4432\cdot10^{-3}=0.$$

Согласно (15.2.9) вычисляем значение χ^2 для проверки нулевой гипотезы: $\chi^2=138{,}33227$. Поскольку $\chi^2>\chi^2_{0{,}05;\,3}=7{,}82$, то гипотезу сферичности данных отвергаем. Далее вычисляем λ_i и K_i: $\lambda_1=2{,}9414048$; $\lambda_2=3{,}4546347\cdot10^{-2}$; $\lambda_3=2{,}4048852\cdot10^{-2}$; $K_1=98{,}046826\%$; $K_2=1{,}1515449\%$; $K_3=8{,}016284\cdot10^{-1}\%$; $K_1+K_2=99{,}198371\%$.

В результате работы программы П 60/34—2 находим коэффициенты для ортонормированных векторов для первой и второй главных компонент:

$$\alpha_1=5{,}7837605\cdot10^{-1};\quad \alpha_2=2{,}692152\cdot10^{-1};$$

$$\beta_1=5{,}7498183\cdot10^{-1};\quad \beta_2=-8{,}0092041\cdot10^{-1};$$

$$\gamma_1=5{,}7868565\cdot10^{-1};\quad \gamma_2=5{,}3483613\cdot10^{-1}.$$

Для каждой из главных компонент имеем $\alpha^2_i+\beta^2_i+\gamma^2_i=1$. Поскольку первая главная компонента имеет $K_1\approx98\%$, то имеет смысл ограничиться только первой главной компонентой.

2. После ввода программы П 60/34—1 и вспомогательных чисел вводим коэффициенты корреляции: 0,2678 П7 0,7426 П8 0,7245 П9. Вычисляем коэффициенты уравнения (15.2.4): В/О С/П (1,8519281=J). Уравнение имеет вид

$$\lambda^3-3\lambda^2+1{,}8519281\cdot\lambda-1{,}400882\cdot10^{-1}=0.$$

Продолжая вычисления, определяем $\lambda_1=2{,}1799515$, $\lambda_2=7{,}3229398\cdot10^{-1}$, $\lambda_3=8{,}715451\cdot10^{-2}$, $K_1=72{,}66505\%$, $K_2=24{,}409799\%$, $K_3=2{,}9251503\%$; $K_1+K_2=97{,}074849\%$. По результатам работы второй программы находим координаты главных компонент (идентифицируйте их самостоятельно):

$$5{,}3514543\cdot10^{-1};\ 7{,}3436558\cdot10^{-1};\ 5{,}2689014\cdot10^{-1};$$
$$-7{,}1707505\cdot10^{-1};\ 6{,}6030759\cdot10^{-1};\ 6{,}9695731\cdot10^{-1}.$$

3. После выполнения первой части программы П 60/34—1 находим коэффициенты уравнения (15.2.4):

$$\lambda^3-3\lambda^2+1{,}6557\cdot\lambda-9{,}5958\cdot10^{-2}=0.$$

Продолжая вычисления по этой программе, находим значения интересующих нас статистик, которые предлагается идентифицировать: $6{,}557794\cdot10^{-2}$; 97,814069; $6{,}368849\cdot10^{-1}$; 76,584573; 2,2975372; 2,1859313; 21,229496; 56,252272.

После идентификации этих статистик сформулируйте свое решение относительно целесообразности применения МГК в данном конкретном случае.

4. Выполнив полностью п. 1—8 инструкции к программе П 60/34, получим следующие числа (идентифицируйте их как те или иные искомые статистики): $3{,}4368746\cdot10^{-1}$; $-8{,}6444381\cdot10^{-1}$; $3{,}669003\cdot10^{-1}$; $6{,}5117448\cdot10^{-1}$; 2,0329217; $3{,}1590381\cdot10^{-1}$; 89,469872; 10,530127; 21,705816; $-4{,}181892\cdot10^{-1}$; 2,171703; 297,29106.

5. Идентифицируйте приведенные ниже числа, полученные по П 60/34, как те или иные статистики: 2,6837; $-7{,}2177\cdot10^{-1}$; 3,9125856; $5{,}1182392\cdot10^{-1}$; $8{,}7319306\cdot10^{-1}$; 1,614983; 29,106435; 53,832766; 82,93920; 17,060797.

Каков результат проверки гипотезы сферичности?

ЗАКЛЮЧЕНИЕ

Жажда знаний, подобно жажде богатств, растет вместе с ее удовлетворением.

Л. Стерн

Применяя в реальных исследованиях те или иные из рассмотренных выше методов и программ, необходимо отчетливо понимать, что «математическая статистика не представляет собой какого-то автомата, в который достаточно заложить статистический материал, чтобы в результате нескольких манипуляций, как на счетной машине, получить готовый результат» [37]. Нередко исследователи испытывают неудовлетворенность от результатов применения тех или иных методов статистического анализа экспериментальных данных. В чем основные причины таких неудачных попыток? Наиболее часто это следствие увлечения формализованной стороной методов при поверхностном знании основных предпосылок их применения и игнорирования имеющихся ограничений. В сочетании с недостаточно ясным и точным представлением о всем многообразии объективно существующих в исследуемом процессе связей и зависимостей это приводит к тому, что, получив отрицательные результаты при использовании одного—двух методов, исследователь не пытается проанализировать возможные причины отрицательного результата. При этом совершается весьма распространенная ошибка: исследователь забывает, что и отрицательный результат также является результатом, содержащим информацию об исследуемом процессе. Только тщательно анализируя как положительные, так и отрицательные результаты применения тех или иных методов анализа экспериментальных данных, можно выбрать оптимальную последовательность все более усложняющихся методов. Четкая ориентация в методах анализа и выборе их последовательности дости-

гается только практикой. В этой части постоянное и систематическое использование ПМК для оперативного анализа экспериментальных наблюдений таит в себе большие возможности. Простота общения с ПМК и его достаточно большие возможности привели к тому, что в настоящее время сфера применения ПМК стремительно расширяется. Общение с ПМК, возможность самому исследователю оперативно оценивать с его помощью результаты исследования на каждом этапе изменяют и саму психологию исследования. Экспериментатор начинает более четко и точно осмысливать и представлять сам эксперимент и его результаты в количественной форме. Овладение разнообразными методами прикладной статистики с их реализацией на ПМК подводит экспериментатора к осознанию возможности и доступности для него решения более сложных задач анализа данных с помощью больших ЭВМ, например ЕС ЭВМ.

В настоящее время имеется немало отечественных и зарубежных пакетов программ статистического анализа для ЕС ЭВМ, предназначенных для их эксплуатации пользователями с различным уровнем подготовки. Так, например, очень удобен в использовании пакет СОМИ (Статистическая Обработка Медицинской Информации), распространяемый Институтом математики АН БССР (г. Минск) и являющийся отечественным аналогом американского пакета BMDP, разработанного специально для медиков и биологов в ВЦ Медицинского центра Калифорнийского университета (г. Лос-Анжелес). Очень много возможностей представляет пользователю и оригинальный пакет ППСА [2, 3], разработанный под руководством С. А. Айвазяна в лаборатории прикладных методов математической статистики и теории вероятностей Центрального экономико-математического института АН СССР (г. Москва). Пакет удостоен «Золотой медали» ВДНХ в 1985 г.

Очень много разнообразных процедур статистического анализа содержат пакеты SAS, ОТЭКС, SPSS и ряд других. Сравнительный анализ этих и многих других пакетов дан в [61].

Описанию подобных программных средств современных методов прикладной статистики посвящена новая серия «Математическое обеспечение прикладной статистики» в издательстве «Финансы и статистика».

ЛИТЕРАТУРА

1. Никитина Е. П., Фрейдлина В. Д., Ярхо А. В. Коллекция определений термина «статистика». — М., 1972. — 45 с.

2. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983. — 471 с.

3. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Исследование зависимостей. — М.: Финансы и статистика, 1985. — 487 с.

4. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.

5. Факторный, дискриминантный и кластерный анализ: Пер. с англ. Под ред. И. С. Енюкова. — М.: Финансы и статистика, 1989. — 215 с.

6. Трохименко Я. К., Любич Ф. Д. Радиотехнические расчеты на микрокалькуляторах. — М.: Радио и связь, 1983. — 256 с.

7. Цветков А. Н., Епанечников В. А. Прикладные программы для микроЭВМ «Электроника Б3-34», «Электроника МК-56», «Электроника МК-54». — М.: Финансы и статистика, 1984. — 175 с.

8. Трохименко Я. К., Любич Ф. Д. Инженерные расчеты на программируемых микрокалькуляторах. — Киев: Техника, 1985. — 328 с.

9. Дьяконов В. П. Расчет нелинейных и импульсных устройств на программируемых микрокалькуляторах: Справочное пособие. — М.: Радио и связь, 1984. — 176 с.

10. Дьяконов В. П. Справочник по расчетам на микрокалькуляторах. — М.: Наука, 1986. — 224 с.

11. Бойко А. Как построить гистограмму. — Наука и жизнь, 1984, № 8, с. 122.

12. Точность производства в машиностроении и приборостроении. — Бородачев Н. А., Абдрашитов Р. М. и др./Под ред. А. Н. Гаврилова. — М.: Машиностроение, 1973. — 567 с.

13. ГОСТ 15895—77. Статистические методы управления качеством продукции. Термины и определения.

14. ГОСТ 16467—70. Статистические показатели точности и стабильности технологических операций. Методы расчета.

15. Бикел П., Доксам К. Математическая статистика. — М.: Финансы, 1983, вып. 1. — 278 с.; вып. 2. — 254 с.

16. Пустыльник Е. И. Статистические методы анализа и обработка наблюдений. — М.: Наука, 1968. — 288 с.

17. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. — М.: Наука, 1983. — 416 с.

18. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. — М.: Финансы и статистика, 1982. — 271 с.

19. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. — М.: Прогресс, 1976. — 495 с.

20. ГОСТ 11. 006—74. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим.

21. Дружинин Н. К. Основные математико-статистические методы в экономических исследованиях. — М.: Статистика, 1968. — 126 с.

22. Поллард Дж. Справочник по вычислительным методам статистики. — М.: Финансы и статистика, 1982. — 344 с.

23. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. — М.: Финансы и статистика, 1983. — 302 с.

24. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. — М.: Статистика, 1973. — 392 с.

25. Химмельблау Д. Анализ процессов статистическими методами. — М.: Мир, 1973. — 957 с.

26. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. — М.: Мир, 1982. — 488 с.

27. Ахназаров С. Л., Кафаров В. В. Оптимизация эксперимента химической технологии. — М.: Высшая школа, 1985. — 327 с.

28. Лукомский Я. И. Теория корреляции и ее применение к анализу производства. — М.: Госстатиздат, 1961. — 375 с.

29. Браунли К. А. Статистическая теория и методология в науке и технике. — М.: Наука, 1977. — 408 с.

30. Болч Б., Хуань К. Дж. Многомерные статистические методы для экономики. — М.: Статистика, 1979. — 317 с.

31. Джонстон Дж. Эконометрические методы. — М.: Статистика, 1980. — 444 с.

32. Андерсон Т. Введение в многомерный статистический анализ. — М.: Физматгиз, 1963. — 500 с.

33. Девис Дж. Статистика и анализ геологических данных.— М.: Мир, 1977. — 572 с.

34. Хемометрика/М. А. Шараф, Д. Л. Иллмэн, Б. Р. Ковальски: Пер. с англ. — Л.: Химия, 1989. — 272 с.

35. Иберла К. Факторный анализ. — М.: Статистика, 1980.— 398 с.

36. Йереском К. Г., Клован Д. И., Реймент Р. А. Геологический факторный анализ. — Л.: Недра, 1980. — 223 с.

37. Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений. — М.: Статистика, 1974.— 240 с.

38. Енюков И. С. Методы, алгоритмы, программы многомерного статистического анализа: Пакет ППСА. — М.: Финансы и статистика, 1986. — 232 с.

39. Митропольский А. К. Техника статистических вычислений. — М.: Наука, 1971. — 576 с.

40. Четыркин Е. М., Калихман И. Л. Вероятность и статистика. — М.: Финансы и статистика, 1982. — 319 с.
41. Леонтьев Н. Л. Техника статистических вычислений. — Л.: Гослесбумиздат, 1961. — 232 с.
42. Сборник задач по математике для вузов. Специальные курсы/Под ред. А. В. Ефимова. — М.: Наука, 1984. — 608 с.
43. Урбах В. Ю. Биометрические методы. Статистическая обработка опытных данных в биологии, сельском хозяйстве и медицине. — М.: Наука, 1964. — 416 с.
44. Снедекор Дж. У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. — М.: Сельхозгиз, 1961. — 503 с.
45. Владимирский Б. М. Математические методы в биологии. — Ростов: Изд-во Рост. ун-та, 1983. — 304 с.
46. Рокицкий П. Ф. Биологическая статистика. — Минск.: Вэйшэйш. школа, 1973. — 320 с.
47. Плохинский Н. А. Биометрия. — М.: Изд-во МГУ, 1970. — 368 с.
48. Меркурьева Е. К. Биометрия в селекции и генетике сельскохозяйственных животных. — М.: Колос, 1970. — 423 с.
49. Меркурьева Е. К. Биометрия в животноводстве. — М.: Колос, 1964. — 311 с.
50. Сепетлиев Д. Статистические методы в научных медицинских исследованиях. — М.: Медицина, 1968. — 419 с.
51. Зайцев Н. Г. Методика биометрических расчетов. — М.: Наука, 1973. — 256 с.
52. Зайцев Н. Г. Математическая статистика в экспериментальной ботанике. — М.: Наука, 1984. — 424 с.
53. Кудрявая К. И. Сборник задач по теории вероятности и математической статистике для гидрометеорологов. Л., 1971. — 84 с.
54. Артемьева Е. Ю. Сборник задач по теории вероятности и математической статистике для психологов. — М.: МГУ, 1969. — 92 с.
55. Шарапов И. П. Применение математической статистики в геологии. — М.: Недра, 1965. — 315 с.
56. Рыжов П. А. Математическая статистика в горном деле. — М., 1965. — 185 с.
57. Батунер Л. М., Позин М. Е. Математические методы в химической технике. — Л.: Химия, 1971. — 824 с.
58. Крастинь О. П. Разработка и интерпретация моделей корреляционных связей в экономике. — Рига.: Зинатне, 1983. — 302 с.
59. Аренс Х., Лёйтер Ю. Многомерный дисперсионный анализ. — М.: Финансы и статистика, 1985. — 230 с.
60. Хальд А. Математическая статистика с техническими приложениями. — М.: ИЛ, 1956. — 664 с.
61. Сильвестров Д. С. Программное обеспечение прикладной статистики: Обзор состояния. Тенденции развития. — М.: Финансы и статистика, 1988. — 240 с.
62. Епанечников В. А., Цветков А. Н. Справочник по прикладным программам для микрокалькуляторов. — М.: Финансы и статистика, 1988. — 320 с.

ПРИЛОЖЕНИЕ

ПЕРЕЧЕНЬ ГОСТОВ ПО ПРИКЛАДНОЙ СТАТИСТИКЕ

№ п/п	ГОСТ	Название
1.	11.001—73	Прикладная статистика. Ряды предпочтительных численных значений статистических характеристик
2.	11.002—73	Прикладная статистика. Правила оценки аномальности результатов наблюдений
3.	11.003—73	Прикладная статистика. Равномерно распределенные случайные числа
4.	11.004—73	Прикладная статистика. Правила определения оценок и доверительных границ для параметров нормального распределения
5.	11.005—74	Прикладная статистика. Правила определения оценок и доверительных границ для параметров экспоненциального распределения и распределения Пуассона
6.	11.006—74	Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим
7.	11.007—75	Прикладная статистика. Правила определения оценок и доверительных границ для параметров распределения Вейбулла
8.	11.008—75	Прикладная статистика. Правила построения и применения вероятностных сеток

ОГЛАВЛЕНИЕ

Предисловие 3
Глава 1. **Развитие прикладной статистики и ее роль в обработке экспериментальных данных**
1.1. От государствоведения до ... статистики 7
1.2. Предмет прикладной статистики 8
Глава 2. **Расчеты на программируемых микрокалькуляторах**
2.1. Редактирование программы 10
2.2. Анализ и изучение особенностей работы программы. 12
Глава 3. **Описательные статистики выборочных массивов. данных**
3.1. Понятие о генеральной совокупности, выборке, выборочных оценках и их свойствах 19
3.2. Построение гистограмм с помощью ПМК . . . 29
3.3. Оценка основных параметров выборки (выборочных характеристик) 30
3.4. Оценка доверительных и толерантных интервалов . 41
Глава 4. **Проверка статистических гипотез**
4.1. Понятие о статистических гипотезах 58
4.2. Проверка статистических гипотез о равенстве дисперсий 62
4.3. Проверка статистических гипотез о равенстве средних 67
4.4. Статистическая проверка гипотез о законах распределения 76
4.5. Проверка статистических гипотез методом последовательного анализа 87
Глава 5. **Анализ статистической связи между качественными переменными**
5.1. Основные типы шкал измерения 95
5.2. Ранговая корреляция Спирмэна 96
5.3. Таблицы сопряженности признаков и коэффициент ассоциации фи (φ) 99
5.4. Критерий случайной последовательности для дихотомической переменной 102
5.5. Точечно-бисериальный коэффициент корреляции Пирсона 103
Глава 6. **Основы дисперсионного анализа наблюдений**
6.1. Основные идеи дисперсионного анализа 108
6.2. Однофакторный дисперсионный анализ 110
6.3. Двухфакторный дисперсионный анализ . . . 122

Глава 7. **Анализ связи между количественными переменными (корреляционный анализ)**
7.1. Задачи корреляционного и регрессионного анализа. 136
7.2. Линейный корреляционный анализ. 140
Глава 8. **Исследование вида зависимости между количественными переменными (линейный регрессионный анализ)** .
8.1. Точность оценки уравнения регрессии 144
8.2. Анализ остатков и последовательностей 150
Глава 9. **Нелинейный корреляционно-регрессионный анализ**
9.1. Нелинейная корреляция 160
9.2. Аппроксимация нелинейной зависимости экспонентой 167
9.3. Аппроксимация зависимости степенным уравнением 169
Глава 10. **Двухмерное нормальное распределение**
10.1. Корреляционный эллипс и двухмерная доверительная область. 182
10.2. Построение корреляционного эллипса 187
Глава 11. **Множественный линейный корреляционно-регрессионный анализ**
11.1. Основы анализа в матричном виде 194
11.2. Точность и значимость оценок в уравнении регрессии 198
11.3. Мультиколлинеарность входных переменных . . 199
11.4. Бета-коэффициенты и ранжирование входных переменных по их влиянию 202
11.5. Полный корреляционно-регрессионный анализ данных с двумя независимыми переменными 204
11.6. Уравнения регрессии с числом входных переменных более двух 206
Глава 12. **Основы ковариационного анализа**
12.1. Основные цели и возможности ковариационного анализа 212
12.2. Программа ковариационного анализа данных с одной сопутствующей переменной 218
Глава 13. **Проверка гипотез о векторах средних**
13.1. Критерий T^2 — Хотеллинга 225
13.2. Проверка статистической гипотезы о равенстве вектора выборочных средних заданному вектору. . 228
13.3. Проверка статистической гипотезы о равенстве двух векторов выборочных средних 234
Глава 14. **Исследование наблюдений методом дискриминантного анализа**
14.1. Понятие о дискриминантном анализе 245
14.2. Линейный дискриминантный анализ двухмерных наблюдений 246
Глава 15. **Применение метода главных компонент (МГК) для анализа наблюдений**
15.1. Основные особенности МГК 252
15.2. Оценка главных компонент 254
15.3. Применение МГК в регрессионном анализе . . . 258
Заключение 263
Литература 266
Библиотека программ 269
Приложение. Перечень ГОСТов по прикладной статистике 375