А. Афифи
С. Эйзен

Статистический анализ

Подход с использованием ЭВМ

Статистический анализ

Подход с использованием ЭВМ

Statistical Analysis
A Computer Oriented Approach
Second Edition

A. A. Afifi
University of California, Los Angeles

S. P. Azen
University of Southern California, Los Angeles

Academic Press
New York·San Francisco·London
1979
A Subsidiary of Harcourt Brace Jovanovich, Publishers

А. Афифи,
С. Эйзен

Статистический анализ

Подход с использованием ЭВМ

Перевод с английского
И. С. Енюкова и
И. Д. Новикова
под редакцией
Г. П. Башарина

Москва «Мир» 1982

УДК 519.24+61

Афифи А., Эйзен С.

Статистический анализ: Подход с использованием ЭВМ. Пер. с англ. — М.: Мир, 1982. — 488 с., ил.

Монография американских ученых, рассчитанная на читателей, знакомых с основами математической статистики, но не имеющих опыта работы с ЭВМ и не знающих программирования. Изложение ориентировано на применение пакетов прикладных программ, приведены примеры из биологии, медицины, гуманитарных наук.

Для математиков-прикладников, научных работников, использующих статистический анализ, для аспирантов и студентов университетов.

Редакция литературы по математическим наукам

Афифи А., Эйзен С.

СТАТИСТИЧЕСКИЙ АНАЛИЗ
ПОДХОД С ИСПОЛЬЗОВАНИЕМ ЭВМ

Научный редактор И. А. Маховая
Мл. научные редакторы И. С. Герасимова, Л. В. Бекренева
Художник А. А. Медников
Художественный редактор В. И. Шаповалов
Технический редактор М. А. Страшнова
Корректор С. А. Денисова

ИБ № 2953

Сдано в набор 11.01.82. Подписано к печати 30.08.82. Формат 60×90¹/₁₆. Бумага типографская № 2. Гарнитура литературная. Печать высокая. Усл. печ. л. 30,50. Усл. кр.-отт. 30,50. Уч.-изд. л. 28,51. Изд. № 1/1730. Тираж 15 000 экз. Зак. № 70. Цена 2 р. 20 к.

ИЗДАТЕЛЬСТВО «МИР», 129820, Москва, И-110, ГСП, 1-й Рижский пер., 2.

Ленинградская типография № 6 ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 193144, г. Ленинград, ул. Моисеенко, 10.

А 1702060000—020 / 041 (01)—82 19—82, ч. 1

От редактора перевода

Вниманию советского читателя предлагается перевод 2-го издания книги, написанной известными специалистами в области статистического анализа с использованием ЭВМ.

Книга имеет характер учебно-справочного руководства и рассчитана на широкий круг читателей с разной математической подготовкой, в том числе и на тех, кто знаком лишь с начальным курсом основ теории вероятностей и математической статистики, еще не имеет опыта работы с ЭВМ и не знает языков программирования. От других книг по прикладной математической статистике данную книгу отличает элементарность и наглядность изложения. Этому способствует большое число хорошо подобранных примеров, занимающих до половины объема книги и не только имеющих иллюстративное назначение, но и представляющих самостоятельный интерес. Поскольку примеры носят в основном биомедицинский характер, книга окажется особенно интересной для медиков, биологов и социологов, начиная со студентов старших курсов. Вместе с тем книга представляет интерес и для лиц, применяющих математическую статистику в других областях. Последние могут рассматривать многочисленные примеры просто как хорошие иллюстрации общих методов статистического анализа.

Другая отличительная особенность книги — ее ориентация на широкое использование ЭВМ и пакетов статистических программ (ПСП). Такая позиция позволила авторам опустить утомительное описание вычислительных процедур, коль скоро они содержатся в ряде имеющихся книг и в документации к ПСП. Сведения о 12 зарубежных ПСП приводятся в гл. 1. Информация об имеющемся в СССР алгоритмическом и программном обеспечении прикладного статистического анализа, а также о ПСП и организациях-разработчиках содержится в публикациях раздела А литературы на русском языке, добавленной редактором перевода. Авторы книги умело оперируют примерами для демонстрации наилучших способов использования программного обеспечения:

выбор самой подходящей для целей исследования программы, использование простых программ для сложного анализа, интерпретация вывода типовых программ и т. д.

Принятый в книге «компьютерный подход» к статистическому анализу оправдывается не только быстрым расширением парка ЭВМ и развитием их математического обеспечения, но и тем, что выполнение многих реальных статистических процедур без ЭВМ просто невозможно. Хочется надеяться, что выход в свет этой книги послужит популяризации современных методов статистического анализа, и в частности ускорению развития, освоения и применения ПСП в многочисленных организациях, приобщившихся за последние годы к использованию ЭВМ при статистической обработке экспериментальных данных.

Естественно, что в такой большой книге не все одинаково удалось. Наглядный стиль изложения, избранный авторами, неизбежно привел к тому, что ряд мест книги оказался излишне описательным. Это прежде всего относится к приложению I, посвященному теоретико-вероятностным основам. В связи с этим в разделе Б литературы, добавленной редактором перевода, содержится список учебников и учебных пособий по теории вероятностей и математической статистике, а в разделе В — небольшой список книг по статистическому анализу; это может облегчить читателю поиск дополнительной литературы на русском языке.

При переводе книги переводчикам и редактору пришлось преодолеть трудности, связанные с разнообразием и неоднородностью терминологии в охваченных книгой областях. В частности, было решено сохранить многочисленные и часто встречающиеся в примерах английские медицинские аббревиатуры, добавив к переводу их перечень с расшифровкой.

Г. П. Башарин

Посвящается
Мэтью Д. и Неннн

Памяти моего отца

Предисловие ко второму изданию

Работая над вторым изданием, мы постарались расширить содержание книги, дополнительно включив в нее современные методы и процедуры анализа данных. С этой целью были добавлены следующие разделы: проверка наборов данных при помощи пакетов статистических программ, робастные оценки параметров («винзоризованные» и M-оценки), обработка отсутствующих наблюдений в многомерном случае, недавно разработанные меры связи в таблицах сопряженности признаков (меры Гудмена — Крускала, коэффициенты ранговой корреляции) и многомерный дисперсионный анализ. Кроме того, мы пересмотрели и добавили много примеров применения математической статистики, почерпнутых из наших исследований в области медицинских приложений (мониторная система наблюдения, применение байесовского метода для многофакторного прогнозирования, применение факторного анализа при разработке карты скрининга нарушений функции легких и т. д.). Кроме того, были включены некоторые классические примеры из медицинской литературы, например фрамингхэмское обследование.

Другие изменения пришлось внести из-за быстрого развития пакетов статистических программ (ПСП). Во втором издании описываются особенности последних версий пакетов BMD-P, SPSS и SAS, а также обсуждаются пакеты GLIM и MINITAB. В книге воспроизводятся выдачи программ из некоторых ПСП.

Наконец, к двум большим наборам данных (наборы A и B) были добавлены несколько меньших. Читатель может использовать многие из представленных результатов вычислений при оценке вновь разрабатываемых статистических программ.

Надеемся, что благодаря этим изменениям второе издание будет лучше отвечать своему назначению — как учебника, так и справочника.

Мы хотели бы поблагодарить Маделин Брадвиг, Лорин Декерт, Жанин Формен, Сару Шонтген, Гейл Уильямс и Джен Уилсон с медицинского факультета (Dept. of Community and Family

Medicine) Университета Южной Калифорнии за большую помощь при подготовке второго издания. Мы благодарим также г-жу Розу Хендерсон за подготовку окончательного варианта рукописи.

Эта работа проводилась при частичной финансовой поддержке Центра биомедицинских исследований (grant NIH BM23732-01). Некоторые примеры в тексте отражают исследования, выполненные в этом Центре.

Вена, Австрия
1977

Предисловие к первому изданию

Когда читатель открывает книгу по статистике, его прежде всего интересует: 1) каков уровень книги, 2) каково ее содержание, 3) отличается ли она от множества других имеющихся в его распоряжении книг по статистике 4) и, наконец, как пользоваться книгой. Вот ответы на эти вопросы.

1. *Уровень книги.* Эта книга написана для читателей, прослушавших только элементарный курс основ теории статистических выводов и не имеющих опыта работы с ЭВМ. В приложении I приводится обзор основных понятий теории статистических выводов, а в гл. 1 читатель познакомится с программистской терминологией и методами, используемыми в книге. Минимально необходимый уровень математической подготовки соответствует курсу, изучаемому в колледжах. Когда мы рассматриваем понятия, требующие математического аппарата, выходящего за рамки этого курса, мы немедленно разъясняем, зачем они нужны и как ими пользоваться. Кроме того, в книге имеются помеченные звездочками разделы, из которых читатель с более основательной математической подготовкой сможет извлечь дополнительные подробности.

2. *Содержание книги.* В книге содержатся как элементарные, так и более сложные разделы. Читатель найдет в ней обзор вероятностных оснований математической статистики и стандартные процедуры статистических выводов. Кроме того, в книгу включены регрессионный и корреляционный анализ, дисперсионный анализ и многомерные методы. Чтобы охватить столь широкий материал, мы исключили математические доказательства и вычислительные формулы и сосредоточили все свое внимание на главном — как применять статистические методы и как интерпретировать полученные результаты.

3. *Отличительные особенности книги.* а) Предполагалось, что все вычисления будут проводиться на ЭВМ. Это позволило нам избежать скучных вычислительных подробностей, которыми обычно изобилуют стандартные учебники, а также рассмотреть

методы регрессионного анализа и пошагового дискриминантного анализа, изложение которых до сих пор было возможно только на гораздо более высоком математическом уровне.

b) Многие сложные вопросы поясняются как математическими формулами, так и словесными комментариями. Вводимые понятия поясняются примерами, почерпнутыми из реальной практики.

c) Показано, как использовать простые программы для сложного анализа. Например, объясняется, как решить задачу простой линейной регрессии, используя дескриптивные программы (описания данных), входящие в пакеты.

d) Разъясняется, как использовать пакеты программ для анализа данных, например для поиска замены переменных, приводящей к нормальному распределению, исследованию остатков для проверки предположений модели и т. д.

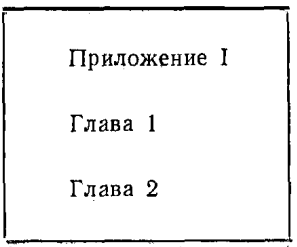
e) Разъясняются также нестандартные способы применения программ из ПСП. Например, показано, как проанализировать план латинских квадратов при помощи факторных программ дисперсионного анализа. Показано также, как проверить линейность регрессионной модели при помощи программ описания данных.

f) Разбросанные по тексту замечания содержат важную дополнительную информацию.

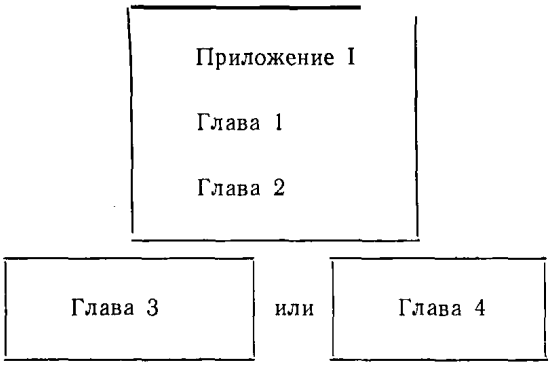
4. *Использование книги.* Книга задумана как справочник по математической статистике для исследователей, в особенности для тех, кто использует пакеты (статистических) программ. Она служит дополнением к сопровождающим пакеты руководствам, поскольку эти руководства обычно описывают только технику работы с программами, т. е. инструктируют, как организовать ввод данных, чтобы получить заданный результат.

Книгу можно использовать как учебное пособие для различных курсов. На следующих диаграммах представлены четыре варианта, соответствующие различным уровням подготовки слушателей.

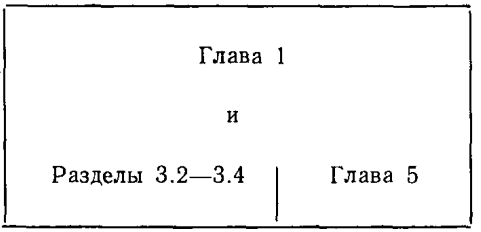
Курс 1 Элементарный прикладной статистический анализ (1 семестр, младшие курсы)



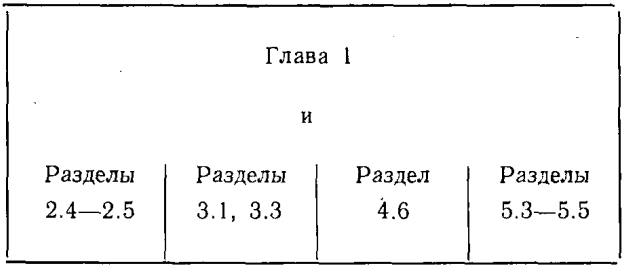
Курс 2 Прикладной статистический анализ (1 год, 3-й курс)



Курс 3 Прикладной многомерный анализ (1 семестр, 4-й курс)



Курс 4 Интенсивный курс обработки данных (1 неделя, 8 часов в день)



В книге принята следующая система нумерации: разделы нумеруются последовательно внутри глав, подразделы, соотношения, замечания, рисунки и таблицы — последовательно внутри разделов:

Разд. i.j обозначает раздел j главы i.

Разд. i.j.k обозначает подраздел k раздела i.j.

Соотношение (i.j.k) обозначает соотношение k раздела i.j.

Табл. i.j.k обозначает таблицу k раздела i.j.
Рис. i.j.k обозначает рисунок k раздела i.j.
Замечание i.j.k обозначает замечание k раздела i.j.
Ссылка в тексте на замечание i.j.k.h обозначает пункт h в замечании i.j.k. Наконец, замечания, помеченные знаком ★, доступны только читателям с более основательной математической подготовкой и могут быть пропущены без ущерба для понимания основного текста.

Лос-Анджелес, Калифорния, 1971

БЛАГОДАРНОСТИ

Мы хотим особо поблагодарить наших студентов Такамуру Асикагу, Энтони Аурьемма, Стьюарта Била, Чарли Бревермана, Икбала Фахми, Томаса Фарвера, Рональда Хасса, Винни Ли, Джоджа Мейера, Сьюзан Сакс и Джирму Вольд-Цадик за их замечания, критику и помощь в проведении многих вычислений, вошедших в книгу. Мы благодарим также Ширли Эйзен и Коллин Гиллен Эйзен за их редакторскую и техническую помощь.

Мы признательны профессору Ричарду Беллману из Университета Южной Калифорнии за его интерес, поддержку и советы по отбору материала для книги. Наша признательность адресована также Вирджинии Зойтл и Лиону Повандру, без административных талантов которых трудно было бы собрать все воедино.

Мы искренне благодарим также замечательных машинисток, которые непостижимым образом переводили наши закорючки в реальные слова — Энн Эйземан, Бетти Хорват, Кэй Ислейб, Джорджи Лам, Джин Рот, Кэти Скофильд и Шэри Уилкокс.

Большинство данных, использованных в примерах, почерпнуто из совместных работ А. Афифи с отделом исследования шока Университета Южной Калифорнии. Понимание нюансов в данных, обсуждаемых в книге, во многом явилось следствием наших обсуждений и совместной работы с докторами Максом Вейлем и Гербертом Шубином из этого отдела. Им мы выражаем свою особую признательность.

Кроме того, мы рады поблагодарить Норму Пэлли и Дэвида Эрбека из того же отдела за важные обсуждения разделов книги, относящихся к обработке данных, а также профессора Вирджинию Кларк из Калифорнийского университета в Лос-Анджелесе за другие полезные дискуссии.

Данные примера 1.4.2 и многих задач в тексте книги помещены с любезного разрешения доктора Джона Чепмена и госпожи Энн Каулсон из того же университета.

Мы обязаны распорядителям литературного наследства сэра Рональда А. Фишера доктору Франку Йэйтсу и издательству «Оливер и Бойд» (Эдинбург) за разрешение перепечатать таблицу III книги «Statistical Tables for Biological Agricultural and Medical Research».

Мы благодарим также сотрудников и редакторов издательства «Академик Пресс» за их помощь, редакционные замечания и т. д.

Помощь в проведении вычислительных работ нам оказывали сотрудники Калифорнийского университета (в соответствии с grant RR-3 от NIH Special Research Resources). Эта работа частично субсидировалась National Institutes of Health Grant No. GM 16197-03, United States Public Health Service research grants HE05570 и GM16462 от National Heart Institute, а также grant HS00238 от National Center for Health Services Research and Development.

1

Введение в анализ данных

Как уже сказано в предисловии, эта книга служит двум основным целям. Первая — описать практику применения основных методов классического статистического анализа как в одномерном, так и в многомерном случаях. Вторая — проиллюстрировать наиболее эффективное использование пакетов статистических программ (ПСП), т. е. показать a) как выбрать наилучшую для целей проводимого анализа программу; b) как интерпретировать различные дополнительные возможности, предоставляемые программой пакета; c) как толковать выход типовой программы и d) как использовать простые программы для сложного анализа.

В этой главе мы рассмотрим предварительные определения и понятия, которые обычно не приводятся в книгах или курсах по статистике. Так, в разд. 1.1 определим виды данных и типы измерений, которые возникают в приложениях, а также опишем элементарные средства для статистических измерений.

В разд. 1.2—1.3 мы изложим общую терминологию, относящуюся к применениям ЭВМ. Так, в разд. 1.2 опишем основные компоненты *вычислительной аппаратуры*, а в разд. 1.3 — основные компоненты *программного обеспечения*, необходимые для понимания принципов составления пакетов программ. В разд. 1.3 приведем также перечень наиболее часто используемых ПСП.

В разд. 1.4 мы опишем подготовку данных для программ из ПСП и обсудим *бланки для программирования* и *операторы формата*. В этом разделе приводятся также два набора данных, которые будут использоваться на протяжении всей книги в примерах и/или упражнениях. В разд. 1.5 обсуждаются требования к хорошему ПСП, а в разд. 1.6 описываются другие важные способы использования ЭВМ для нужд статистики. В разд. 1.7 описываются этапы проверки данных, предшествующие дальнейшему статистическому анализу.

1.1. Данные, измерения и вычислительные средства

Термин *данные* весьма популярен в научных исследованиях. В широком смысле он означает фактический материал, являющийся основой для обсуждения или принятия решений, а в статистике — это информация, пригодная для анализа и интерпретации. Действительно, некоторые статистики рассматривают статистический анализ как анализ данных (Tukey (1962)). В этой книге *наблюдения* служат реализацией некоторой случайной величины, и они поставляют данные для изучаемой проблемы. Следовательно, термины «данные», «наблюдения» и «реализации» являются синонимами и могут заменять друг друга.

В настоящем разделе мы обсудим типы данных, возникающих в научных исследованиях. Данные получаются в результате *измерений* индивидуумов или подопытных образцов из исследуемой популяции. Под измерением мы понимаем присвоение *символов* подопытным образцам в соответствии с некоторым правилом. Эти символы могут быть буквенными и представлять *классы* или *категории* в популяции или числовыми. Числовые символы также могут представлять категории в популяции или быть числами. В первом случае к ним нельзя применять правила арифметики, во втором — можно. Например, если 1 обозначает класс мужчин, а 2 — женщин, то в этом контексте $1 + 2$ не имеет смысла. Однако если 1 — число долларов, заработанных за некоторый день, а 2 — за следующий день, то $1 + 2 = 3$ имеет смысл и означает, что за два дня заработано 3 доллара.

Шкала и единицы измерений могут быть самыми разными. Например, для любого индивидуума из популяции взрослых в США мы можем измерить a) пол; b) социальное положение; c) температуру; d) рост. Очевидно, что шкалы этих четырех измерений совершенно различны по существу, так как в a) можно сказать, что пол одного индивидуума *отличен от* пола другого; в b) можно сказать, что положение одного отличается и *выше*, чем у другого; в c) можно сказать, что температура одного отлична, выше и *на сколько выше*, чем у другого; в d) можно сказать, что рост одного отличен, больше, на сколько больше и *во сколько раз больше*, чем у другого. Эти четыре примера представляют четыре типа шкалы измерений, предложенные С. С. Стивенсом (Churchman, Ratoosh (1959), гл. 2) и получившие следующие названия: *шкала наименований*, *порядковая шкала*, *интервальная шкала* и *шкала отношений*. Обсудим теперь коротко каждую из шкал.

1. *Шкала наименований.* Эта шкала используется только для классификации индивидуумов в популяции. Каждому классу присваивается свое обозначение так, чтобы обозначения различных классов не совпадали. Например, если индивидуумы класси-

фицируются по полу, то двум классам можно присваивать соответственно буквы M и F, слова MALE и FEMALE или цифры 1 и 2.

Структура шкалы наименований не изменяется, если произвести взаимно однозначную подстановку обозначений. Так, в приведенном выше примере можно подставить 1 вместо M и 2 вместо F, или 2 вместо M и 1 вместо F, или 100 вместо M и 1000 вместо F и т. д.

Повторим, что арифметические операции не имеют смысла для шкалы наименований. Поэтому ни *медиана*, ни *среднее* не имеют смысла. Подходящей статистикой положения центра (центральной тенденции) является *мода*, так как она не изменяется при взаимно однозначной подстановке обозначений. Например, если мужчин больше, чем женщин, то мода описывает класс «мужчины» независимо от того, будет ли он обозначен через M, 1, 2 или 1000.

2. *Порядковая шкала.* Эта шкала позволяет не только разбивать индивидуумы на классы, но и упорядочить сами классы. Каждому классу мы присваиваем различные обозначения так, чтобы порядок обозначений соответствовал порядку классов. Если мы нумеруем классы, то классы находятся в числовом порядке; если обозначаем классы посредством букв, то классы находятся в алфавитном порядке; если обозначаем классы словами, то порядок соответствует смыслу слов. Пусть, например, мы хотим классифицировать индивидуумы по трем социально-экономическим категориям — низкий, средний, высокий. Если мы решили упорядочить эти классы от низкого к высокому, то можем присвоить им такие обозначения: 1 — низкий; 2 — средний, 3 — высокий, или X — низкий, Y — средний, Z — высокий, или НИЗКИЙ, СРЕДНИЙ, ВЫСОКИЙ. С другой стороны, мы можем упорядочить классы сверху вниз, приняв, что 1 — высокий, 2 — средний, 3 — низкий и т. д. В этом примере цифры и буквы являются последовательными, но это не обязательно, так как можно обозначить, например, 1 — низкий, 10 — средний, 100 — высокий, или A — низкий, P — средний, Z — высокий и т. д.

Структура порядковой шкалы сохраняется при любой взаимно однозначной подстановке, которая сохраняет порядок. Например, $1 \to 2$, $2 \to 3$, $3 \to x$, где $x > 3$ — допустимая перестановка, а $1 \to 2$, $2 \to 3$, $3 \to 1$ — недопустимая.

Арифметические операции для этой шкалы также не имеют смысла, так что подходящие статистики положения должны не зависеть от значения наименований классов. Поэтому медиана и мода являются подходящими мерами положения центра.

3. *Интервальная шкала.* Эта шкала позволяет не только классифицировать и упорядочивать индивидуумы, но и количественно