

Афифи А., Эйзен С.

Статистический анализ.
Подход с использованием ЭВМ

Москва
«Книга по Требованию»

УДК 62-63
ББК 30.6
А94

А94 **Афифи А.**
Статистический анализ.: Подход с использованием ЭВМ / Афифи А., Эйзен С. – М.: Книга по Требованию, 2013. – 488 с.

ISBN 978-5-458-27364-0

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.

ISBN 978-5-458-27364-0

© Издание на русском языке, оформление
«YOYO Media», 2013

© Издание на русском языке, оцифровка,
«Книга по Требованию», 2013

Эта книга является репринтом оригинала, который мы создали специально для Вас, используя запатентованные технологии производства репринтных книг и печати по требованию.

Сначала мы отсканировали каждую страницу оригинала этой редкой книги на профессиональном оборудовании. Затем с помощью специально разработанных программ мы произвели очистку изображения от пятен, клякс, перегибов и попытались отбелить и выровнять каждую страницу книги. К сожалению, некоторые страницы нельзя вернуть в изначальное состояние, и если их было трудно читать в оригинале, то даже при цифровой реставрации их невозможно улучшить.

Разумеется, автоматизированная программная обработка репринтных книг – не самое лучшее решение для восстановления текста в его первозданном виде, однако, наша цель – вернуть читателю точную копию книги, которой может быть несколько веков.

Поэтому мы предупреждаем о возможных погрешностях восстановленного репринтного издания. В издании могут отсутствовать одна или несколько страниц текста, могут встретиться невыводимые пятна и кляксы, надписи на полях или подчеркивания в тексте, нечитаемые фрагменты текста или загибы страниц. Покупать или не покупать подобные издания – решать Вам, мы же делаем все возможное, чтобы редкие и ценные книги, еще недавно утраченные и несправедливо забытые, вновь стали доступными для всех читателей.



Серия Книжный Ренессанс

www.samizday.ru/reprint

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 Ненни
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 Бравдиг, 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.

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

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

д) 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.

е) 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.

ф) 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.

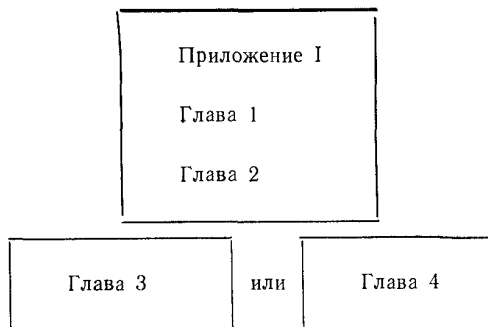
Курс 1 Элементарный прикладной статистический анализ
(1 семестр, младшие курсы)

Приложение I

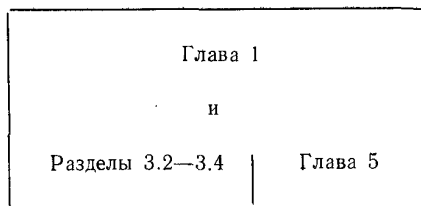
Глава 1

Глава 2

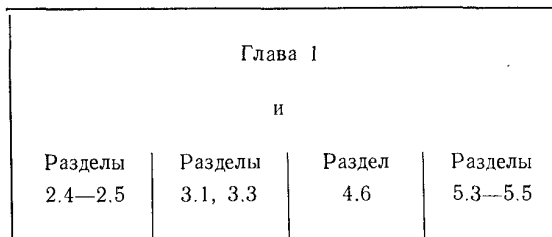
Курс 2 Прикладной статистический анализ
(1 год, 3-й курс)



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 Эйзман, Бетти Хорват, Кэй Ислейб, Джорджи Лам, Джин Рот, Кэти Скофильд и Шэри Уилкоккс.

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 а) как выбрать наилучшую для целей проводимого анализа программу; б) как интерпретировать различные дополнительные возможности, предоставляемые программой пакета; в) как толковать выход типовой программы и д) 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.