Содержание

Предисловие к четвертому изданию на русском языке

Статистика — описание реальных фактов,
в особенности относящихся к
современной жизни народа.
W. Roscher

Статистике можно доверять лишь тогда,
когда вы сами ею манипулируете.
Американское изречение

«Наглядная медицинская статистика» — книга весьма полезная для медиков, которые проводят свои исследования по базам данных, а также для обучения студентов. В 4-м издании в некоторые главы помещены новые тексты и примеры использования конкретных методов статистического анализа, а также добавлено еще одно приложение.

В каждой главе авторы приводят цели изучения методов статистического анализа, систематизируют полученный в результате обучения студентов и собственных исследований опыт. Редактор русского издания также имеет немалый опыт в данной сфере.

Такая работа давно была начата мною — еще во время работы в НИИ, когда успешно применил методы статистического анализа в процессе написания кандидатской диссертации; затем в 1990 г. опубликована первая книга по статистическому анализу под моим авторством; в 1997 г. мною создан сайт Биометрика[1] (URL: http://www.biometrica-tomsk.ru/), который содержит много информации по статистическому анализу. В 1999 г. в московской газете «Поиск» (№ 20) была опубликована моя статья «В новый век — с доказательной биомедициной»[2] (URL: http://www.biometrica-tomsk.ru/poisk.htm). В ней и было предложено Высшей аттестационной комиссии РФ создать свой сайт и требовать публикации диссертаций в Интернете за три месяца до их защиты, чтобы эти публикации обеспечили и оценку таких диссертаций, и доступ к полезной информации для всех российских исследователей.

После перевода в Томский государственный университет преподавал статистику студентам и группам преподавателей, а также на выездных семинарах (URL: http://www.biometrica-tomsk.ru/biostat_5.htm). За последние 40 лет мною выполнены более 1500 работ по статистическому анализу баз данных, составляемых медиками, биологами, социологами, техниками и другими специалистами, которые создавали диссертации, книги или писали статьи. Многие эти работы в виде статей и докладов на международных конференциях по доказательной медицине выложены на сайте Биометрика.

Исследовательская работа в области медстатистики ведется мною уже более 15 лет, а результаты представлены на сайте Биометрика — имеются примеры положительных результатов такого анализа, полученных с нашей помощью, а также немало и примеров весьма отрицательного использования и описания статистического анализа в статьях и диссертациях. И вот этот многолетний опыт использования продуктивного статистического анализа генерирует для читателей этой книги полезные направления использования этого издания.

При чтении этой книги необходимо «смотреть» и «видеть». Смотреть нужно практически все, а вот видеть результаты чтения — гораздо больше. Если же не видеть большой пользы от многих методов статистического анализа, то результаты чтения или исследования будут примитивными. Вот почему авторы данной книги и сообщают о том, что для здравоохранения и медицинской науки будет польза от описанных методов. Чтобы объяснить высокую практическую ценность данной книги для медиков, рассмотрим вначале возникновение в мире самого понятия статистики, а затем и динамику развития и использования медицинской статистики в России.

Первоначально статистика в медицине использовалась лишь в формате уточнения уровня смертности населения в конкретной стране. Например, такая оценка производилась в Англии уже в XIV веке. Подобная же описательная статистика была представлена в 1614 г. в книге «О статистической медицине», автор которой Санторио, президент венецианской Коллегии врачей. Основным методом такой статистики являлось лишь описание в медицине количества и доли пациентов, методов их лечения, разных результатов, использования лекарств и т.п. Подобная же статья была опубликована далее и во Франции. «*В 1835 г. уролог Ж. Сивиаль опубликовал статью, из которой следовало, что после бескровного удаления камней мочевого пузыря выживают 97% больных, а после 5175 традиционных операций выжили только 78% больных*»[1].

То есть такая статистика, будучи фрагментом математики, позволяла использовать ее практически каждому медику. Польза этого авторского подхода была в том, что данную книгу «О статистической медицине» стали переиздавать во многих иных странах на других языках. А затем новые авторы в своих подобных книгах стали использовать не только табличные показания, доли результатов, их обсуждения, но и описания теории и практики новых

[1] Леонов В.П. Зачем нужна статистика в доказательной медицине? // Армянский медицинский реферативный журнал. 2012. Вып. 9. С. 184–193. URL: http://www.biometrica-tomsk.ru/erevan_3.html

методов статистического анализа. В частности, в XVII–XX в. это издания следующих авторов: Д. Граунта, Я. Бернулли, М. Ломоносова, В. Пэтти, А. Кетле, Ф. Гальтона, Г. Дункера, В.Н. Татищева, Н. Бейли, К. Пирсона, В. Уэлдона, Ч. Дарвина, Н.И. Пирогова, П.Ф. Рокицкого, Д. Химмельблау, В.Ю. Урбаха, В.В. Алпатова и др. Например, бельгийский статистик А. Кетле (1796–1874 гг.) начал использовать статистику в медицине, публикуя эти результаты в своих статьях. О важности такого использования статистики сообщал и российский военный хирург Н.И. Пирогов, который писал: «*...Приложение статистики для определения диагностической важности симптомов и достоинства операций можно рассматривать как важное приобретение новейшей хирургии*»[1].

Отметим, что первоначально Россия в развитии статистики не отставала от передовых европейских стран [222]. Так, еще в 1804 г. при Академии наук впервые был организован факультет статистики, а в гимназиях и училищах уже начали преподавать статистику. В 1806 г. был организован выпуск «Статистического журнала» и многих учебников по статистике, а в 1863 г. в правительстве России был организован «Статистический совет». Признанием успехов российской статистики стало проведение в 1872 г. в Санкт-Петербурге восьмой сессии Международного статистического конгресса. Вторая половина XIX в. и начало XX в. для российской статистики ознаменовались значительным подъемом теоретических и прикладных работ. Доминировала в этом Петербургская математическая школа. Большинство ее представителей работали в Петербургском университете и Петербургском политехническом институте. Ведущий статистик того времени А.А. Чупров писал: «*Будущий историк человеческой мысли, окидывая взором современную нам эпоху конца XIX — начала XX в., отметит как ее характерную черту стремление научного знания облекаться в статистические формы. ...Без преувеличения можно сказать: рост современной науки идет под знаком интереса к массовым явлениям, и скоро не будет той ветви знания, куда бы с большим или меньшим успехом не простирали бы своего влияния статистические формы знания*» [222]. Все это приводило к широкому проникновению статистики в российскую медицину. Активным сторонником использования статистики был и основоположник военно-полевой хирургии Н.И. Пирогов. В своем учебнике по хирургии он написал: «*Я принадлежу к ревностным сторонникам рациональной статистики и верю, что* ***приложение ее к военной хирургии есть несомненный прогресс***» [222]. В российских медицинских журналах стали публиковаться статьи с результатами статистических выводов.

Одним из активных сторонников статистических методов был и петербургский медик М.К. Зенец. В 1874 г. в «Военно-медицинском журнале» он опубликовал статью, в которой писал: «*Медицина есть именно одна из тех областей человеческого ведения, в которой можно ожидать* ***от приложения статистико-математического метода самых плодотворных результатов***» [222]. При этом статистика становилась уже более сложной, чем в первоначальном периоде, и более продуктивной. Это означало, что эти новые, гораздо более сложные методы статистического анализа становились уже доступными не каждому медику. Вот почему уже в XIX в. стали появляться научные специалисты именно по статистике в медицине и биологии, а также и сами математики именно по статистике. А во второй половине XX в. стали появляться в Интернете и сайты по таким направлениям статистики. «*Европейские исследователи уже в начале XX в. создавали лаборатории биостатистики, ориентированные на использование методов статистики в биомедицине. В это же самое время начали издаваться и специализированные журналы данной тематики. Так в Англии в 1901 г. появился такой известный журнал, как «Biometrika», в США в 1945 г. журнал «Biometrics», в 1959 г. в Германии появился журнал «Biometrische Zeitschrift», затем появились «Психометрика», «Технометрика», «Эконометрика» и «Наукометрия». Материалы этих журналов были посвящены применению статистики в различных областях науки. Такие журналы не только выполняют обучающую функцию, но и прививают авторам и читателям вкус и потребность к грамотной статистической обработке экспериментальных данных. Уже 20 лет издательством JOHN WILEY & SONS издается специализированный журнал «Statistics in Medicine», а издательством ELSEVIER выпущено уже более 50 томов журнала «Computer Methods and Programs in Biomedicine». В 1978 г. было организовано Международное общество клинической биостатистики (ISCB), национальные отделения которого есть в нескольких десятках стран, включая США, Англию, Францию, Италию, Канаду, Испанию, Польшу, Венгрию, Южную Африку, Кению и т.д. 33-я ежегодная конференция Международного общества клинической биостатистики состоится 19–23 августа 2012 г. в Бергене (Норвегия). Однако Россия не участвует в работе этой организации. В дореволюционный период в России и далее в первые десятилетия существования СССР статистика широко применялась в медицине [8]. Однако с началом гонений на генетику в СССР «предали анафеме» и статистику, поскольку эта наука была одним из мощнейших инструментов генетики. Этому же способствовало и преследование кибернетики, поскольку* ***статистика является одной из составных частей кибернетики****. Вот какое толкование кибернетики дается в «Кратком философском словаре» [10]: «****Кибернетика — реакционная лженаука, возникшая в США после Второй мировой войны...*** *Кибернетика ярко выражает одну из основных черт буржуазного мировоззрения — его бесчеловечность, стремление превратить трудящихся в придаток машины, в орудие производства и орудие войны». В середине XX в. ВАК СССР отказывал медикам в получении ученых степеней кандидатов и докторов наук по причине использования ими «****буржуазной статистики****» [8]. В дальнейшем в СССР и далее в Российской Федерации извращенная система*

[1] Пирогов Н.И. Об успехах хирургии в течение настоящего пятидесятилетия. Записки по части врачебных наук. Кн. 4. СПб., 1849.

аттестации научных и научно-педагогических работников, реализуемая ВАК, привела к тому, что статистика стала не инструментом качественных исследований в биомедицине, а средством онаучивания безграмотно выполненных в биомедицине экспериментальных исследований. Весьма обстоятельно об этом аспекте системы аттестации рассказано в статье известного специалиста по ДМ Бащинского С.Е. [11]»[1].

Реально все без исключения науки используют эмпирические и теоретические методы познания. В эмпирической составляющей используются наблюдения, измерения и эксперименты. Теоретическая составляющая в различных науках содержит разные комбинации аналитико-синтетических методов. Накопление эмпирических данных во всех науках способствовало возникновению такой науки, как статистика, которая далее переросла в математическую статистику.

До появления математической статистики во всей медицине активно использовался тот же статистический принцип обобщения подобных эмпирических знаний. В итоге результатами этих обобщений были описания числа медицинских методик лечения на словесном (вербальном) уровне. Типичными примерами таких описаний являются труды Гиппократа и Авиценны, то есть в медицине использовалась вербальная статистика. Отметим, что у опытных ученых-медиков описания подобных обобщений на вербальном уровне даются весьма понятным языком. В результате этого такие обобщения становятся доступными и вызывающими доверие у медиков-практиков.

Таким образом, собираемые эмпирические данные в итоге имеют два канала обращения: у медиков-исследователей, обобщающих эти данные, и у медиков-практиков. *«Фактически **эти методики есть словесные алгоритмы увеличения вероятности достижения благоприятных исходов лечения** (нелетальный исход, возврат организма пациента в здоровое состояние и т.п.). Однако такие вербальные алгоритмы не имеют возможности оценки числовых значений вероятности этих благоприятных исходов. При этом в самих методиках используются количественные показатели (дозы лекарственных препаратов, длительность и интенсивность лечебных процедур, значения показателей, полученных при анализе крови, мочи, ЭКГ, и т.д.). Вместе с этими показателями используются во всех методиках и качественные показатели, например, пол, генотип, тип инфекции во время беременности, наличие эндемического зоба, форма комплекса QRS и т.д. Именно сочетание различных видов информации о пациенте, а также информации о самом лечении и позволяет медику определять возможность благоприятного исхода лечения»* [241].

После того как появилась математическая статистика, использующая математический язык, стали развиваться и прикладные разделы статистики. В настоящее время, кроме медицинской статистики, есть аналоги: биометрика, психометрика, эконометрика, социометрия, технометрика, хемометрика, наукометрика и т.д. Соотношение эмпирических и теоретических значений в таких науках определяли соотношения между умением и знанием. «*На протяжении многих веков медицина больше умела, чем знала. В настоящее время знание, научное обоснование, как правило, предшествует умению, т.е. практическому использованию. В перспективе под умение в возрастающей и ускоренной по времени степени будет подводиться знание или адекватная научная, теоретическая основа»* [384].

По мере накопления эмпирических данных стала возрастать и доля теоретической составляющей в каждой из этих наук. В итоге концентрация эмпирических данных приводила к переходу от вербального описания теоретических составляющих к описаниям специальным формальным языком. Такие языки появились в химии, физике, математике, логике и т.д. В XVI–XVII вв. получил развитие математический язык, который более краток, чем обычный, словесный язык. И этот математический язык активно входил и в теорию статистического анализа. Появление этого математического языка способствовало резкому подъему в развитии физики. В СССР и в России физика — одна из самых активно развивающихся наук. Так, с 1958 по 2003 г. Нобелевскую премию получили 10 физиков из СССР и России. Поскольку Нобелевская премия не присуждается за научные достижения в области математики, то в 1936 г. был создан «эквивалент» этой премии — Филдсовская премия, которую часто называют «Нобелевской премией для математиков». За это время премия была вручена 48 математикам, из них 8 лауреатов из СССР и России. Присуждают российским математикам и другую престижную награду — Абелевскую премию. Однако за 120 лет Нобелевская премия по медицине была лишь однажды присуждена в 1904 г. россиянину — физиологу И.П. Павлову. Тогда как в США эту премию получили более 100 медиков, а в Англии, Германии, Франции, Швеции более 50 медиков и т.д. Каковы же причины такого контраста? Одной из основных причин этого и стало удаление в СССР статистических технологий обобщения эмпирической информации из отечественной биологии и медицины. Этот процесс начался в СССР в середине 20-х годов XX в.

После революции в 1917 г. интерес к применению статистики в научных исследованиях в России не уменьшился. Продолжала работу школа статистиков в Петербургском университете, где работал известный медицинский статистик Л.С. Каминский. Он был автором многих известных учебников по медицинской статистике, работал также заведующим кафедрой статистики в Военно-медицинской академии, был экспертом Всемирной организации здравоохранения по санитарной статистике. Однако начиная с 1925–1926 гг. возрастают усилия вла-

[1] Леонов В.П. Зачем нужна статистика в доказательной медицине? // Армянский медицинский реферативный журнал. 2012. Вып. 9. С. 184–193. URL: http://www.biometrica-tomsk.ru/erevan_3.html

сти втянуть в политические распри и научные сферы. Так, в 1926 г. в одной из статей влиятельного в то время журнала «Под знаменем марксизма» было написано: *«...Современное естествознание так же классово, как и философия и искусство... **Оно буржуазно в своих теоретических основаниях**»* [222]. А в 1930 г. в редакционной статье журнала «Естествознание и марксизм» прямо утверждалось, что *«...философия, естественные **и математические науки так же партийны,** как и науки экономические или исторические»* [222]. В это время в учебниках по математике данный лозунг уже провозглашался как реальность. Так, в одном из учебников по статистике было написано*: «...**Статистика, как и всякая другая наука, наука партийная»***. Такое отрицательное отношение к статистике вызывалось тем, что результатами использования статистики были истинные, объективные выводы, которые весьма не устраивали власть. Следствием этого и стало изгнание из статистики математики «как математического формализма».

Математика в статистике в этот период противопоставлялась «правильной марксистской статистике». К концу 30-х годов XX в. статистический центр в Ленинградском университете перестает существовать, ликвидирован был статистический цикл, упразднена статистическая кафедра [222]. Очевидно, что это не могло не сказаться на отношении к статистике и не привести к снижению интереса к ней. В качестве симптоматичного свидетельства этого приведу следующий пример. В 1930 г. был издан перевод зарубежной книги А. Боули «Элементы статистики. Общие элементарные методы» (Ч. 1. М.; Л.: Гос. изд-во, 1930. 299 с.). В 1998 г. автор этого предисловия, работавший доцентом в ТГУ, искал книгу в научной библиотеке ТГУ. И когда нашел ее, то оказалось, что она так и осталась неразрезанной как по горизонтальным, так и по вертикальным сгибам. То есть за прошедшие 68 лет эту книгу ни разу не читали ни сами преподаватели, ни студенты ТГУ.

В тот же период началась борьба и с генетиками, которых власть считала учеными-вредителями. Так, из Москвы был выслан С.С. Четвериков — создатель экспериментальной и популяционной генетики. Это был первый ощутимый удар по медицинской статистике, потому что именно С.С. Четвериков первым в России начал читать в МГУ курс биометрики с основами генетики в 1919 г., а в 1924 г. он читал уже курс «Введение в биометрию». Этот подход к оценке научных исследований наглядно представлен в статье одного из идеологов того времени Э. Кольмана[1] «Вредительство в науке».

По его мнению, основным признаком таких «вредителей» является *«...исключительное обилие вычислений и формул, которыми так и пестрят вредительские работы. ...Математические уравнения сплошь да рядом придают враждебным социалистическому строительству положениям якобы бесстрастный, объективный, точный, неопровержимый характер, скрывая их истинную сущность»* [222]. Это отношение тогдашней власти к статистике весьма наглядно отражено в следующей подборке цитат из публикаций того времени. *«Методы реакционной английской статистики как нельзя лучше подходят к реакционной менделе-моргановской школе в биологии»; «Советские статистические методы являются самыми передовыми, ибо они базируются на гениальных трудах Ленина и Сталина»; «Статистическая теория и наука могут опираться только на философию Маркса–Энгельса–Ленина–Сталина. Диалектический материализм и марксистско-ленинская политическая экономия, а не закон больших чисел являются основой статистики как науки»* [222].

Вершиной борьбы тогдашней власти с биомедицинской наукой стала августовская сессия ВАСХНИЛ 1948 г., завершившаяся разгромом генетики [222].

Подробный перечень того, что вменялось в вину последователям Менделя в медицине, дан в книге В.М. Банщикова «Против реакционных биологических теорий в медицине» (М.: Медгиз, 1948 г. 63 с.). После этого из учебных программ вузов удалили генетику и статистику, а в библиотеках уничтожали книги этой тематики. Результатом гонений за использование статистики в биологии и медицине стало изменение политики ВАК СССР. Так, по причине использования методов статистики в биомедицинских диссертациях стали отказывать диссертантам в присуждении ученых степеней. Один из таких примеров приведен в статье секретаря Фрунзенского райкома ВКП(б) г. Москвы Е. Фурцевой (будущего министра культуры СССР. — *В.Л.*) «Партийное руководство научными учреждениями», опубликованной в газете «Правда» от 3 августа 1949 г.: *«Ученый совет Московского медицинского института утвердил, например, две диссертации — одну на соискание ученой степени кандидата наук (Г.Л. Лемперта), другую — на степень доктора медицинских наук (Г.П. Сальниковой). Авторы некритически использовали данные **лживой, тенденциозной буржуазной статистики** и пришли к чудовищно извращенным, лженаучным выводам. Однако коммунисты — члены ученого совета 1-го Московского медицинского института — прошли мимо лженаучных утверждений «диссертантов» и голосовали за присвоение им ученых степеней. И правильно решила Высшая аттестационная комиссия Министерства высшего образования СССР, отказав Сальниковой и Лемперту в присвоении ученых степеней»* [222].

[1] Эрнст Кольман был с 1929 по 1943 г. членом редколлегии журнала «Под знаменем марксизма». С 1931 г. он возглавлял Институт красной профессуры, а с 1939 по 1945 г. являлся заведующим сектором диалектического материализма Института философии АН СССР, затем возглавлял кафедру высшей математики одного из московских вузов. В 1976 г. ему удалось выехать в Швецию, где он получил политическое убежище, и в том же году он вышел из КПСС, в которой состоял 58 лет. За несколько лет до своей смерти, в 1979 г., он издает мемуары «Мы не должны были так жить», в которых раскаивается в содеянном им. — *Прим. ред.*

Напомним, что долгие годы именно Т.Д. Лысенко был заместителем председателя ВАК СССР. И он весьма отрицательно относился к использованию исследователями статистических методов анализа. Так, в своей статье «По поводу статьи академика А.Н. Колмогорова» (журнал «Доклады Академии наук СССР». 1940. Том. 28, № 9) Т.Д. Лысенко написал следующее: «*Поэтому-то нас, биологов, и не интересуют математические выкладки,* ***подтверждающие практически бесполезные статистические формулы менделистов***».

Известный российский пропагандист биометрии А.А. Любищев, единственный из СССР член Международного биометрического общества, 30 июля 1955 г. закончил статью «**Об аракчеевском режиме в биологии**», в которой написал следующее: «*Так, в результате гонения на математическую статистику, связанную с законами Менделя, из программы преподавания биологии* ***в университетах были совершенно изгнаны высшая математика и вариационная статистика***».

Листая медицинские журналы тех лет, мы не найдем там статистики: медицина оставалась лишь описательной наукой. Это приводило к значительному снижению качества обобщения накапливаемых медицинских данных, уменьшению теоретической составляющей медицинской науки и снижению качества практической медицины. В 60-е годы XX в., после низвержения Т.Д. Лысенко и очевидных успехов прикладной статистики в технике и точных науках, стал вновь возрастать интерес к использованию статистики в биологии и медицине. В журналах «Вопросы философии» и «Вестник высшей школы» периодически стали появляться статьи на эту тему. Так, В.В. Алпатов в статье «О роли математики в медицине» писал: «*Чрезвычайно важна математическая оценка терапевтических воздействий на человека. Новые лечебные мероприятия имеют право заменить собою мероприятия, уже вошедшие в практику, лишь после обоснованных статистических испытаний сравнительного характера. ...****Огромное применение может получить статистическая теория в постановке клинических и внеклинических испытаний*** *новых терапевтических и хирургических мероприятий. ...Здесь необходимо подчеркнуть, что* ***математик-статистик должен включаться в работу медика-экспериментатора на самых начальных этапах этой работы***» [222].

При этом включение профессиональных статистиков в научные исследования по медицине определяется отношением руководителей медицинских организаций, в частности, пониманием продуктивности использования современных методов статистического анализа по их базам данных и уровнем понимания сложности реальных медицинских технологий. При упрощенных уровнях этих деталей руководители медицинских НИИ и медицинских вузов вместо лабораторий биостатистики открывают церкви[1].

Невысокий уровень издаваемых ранее книг по медицинской статистике, примитивный подход к обучению медиков статистике и отсутствие специалистов по биостатистике не позволили СССР и России за прошедшие полвека подняться в использовании медицинской статистики до мирового уровня. Проведенный нами критический анализ тысяч медицинских статей, монографий и диссертаций [220, 226, 229, 235, 236, 238] показал, что статистическая парадигма российской медицины сводится к примитивной сдвиговой гипотезе. Иными словами, подавляющее большинство исследователей в медицине считают, что основное различие между группами сравнения пациентов сводится лишь к сравнению средних значений («температура тела больных выше температуры у здоровых»). Свидетельством этой сдвиговой парадигмы является доминирование в публикациях результатов сравнения групповых средних с помощью *t*-критерия Стьюдента. Причем делают такие сравнения без проверки двух условий корректности использования этого метода, тогда как два этих совместных условия в реальных базах данных вместе имеют положительное состояние лишь порядка 5–10%. Фактически в отечественных публикациях наблюдается рост статистических вампук, иначе говоря, в российской медицинской науке идет процесс статистической вампукизации [236]. Большая часть информации с результатами статистики, публикуемая в настоящее время в журналах и диссертациях, содержит ошибки использования статистических методов анализа. Авторами таких публикаций являются не только начинающие исследователи, но и ученые, являющиеся кандидатами и докторами медицинских наук, академиками, ректорами вузов, директорами медицинских НИИ и т.п. Типичным примером такой публикации является статья, вышедшая в журнале «Генетика»[2], среди авторов которой два академика РАМН, директора НИИ РАМН.

Таким образом, огромное количество собираемых в медицине эмпирических данных не позволяет медикам-исследователям, не привлекающим к своим исследованиям профессиональных статистиков, получать из своих баз данных полезную информацию. Более того, получаемые в результате ложные выводы из-за некорректного использования статистики являются по сути «информационным талидомидом»[3], информационным ядом. При этом подобные

[1] URL: http://socialte.tomsk.ru/xram-prmchc-elisavety-v-nii-kardiologi/

[2] Спиридонова М.Г., Степанов В.А., Пузырев В.П., Карпов Р.С. Анализ взаимосвязи полиморфизма С677Т гена метилентетрагидрофолатредуктазы с клиническими проявлениями атеросклероза // Генетика. 2000. Вып. 9. С. 1269–1273. URL: http://www.biometrica-tomsk.ru/kk/index_3.htm#33

[3] Талидомид — седативное снотворное лекарственное средство, получившее широкую известность из-за своей тератогенности. В период с 1956 по 1962 г. было установлено, что в результате потребления талидомида роженицами порядка 40 000 человек получили периферический неврит, от 8000 до 12 000 новорожденных родились с физическими уродствами, из них лишь около 5000 не погибли в раннем возрасте, оставшись инвалидами на всю жизнь. В результате многие страны пересмотрели практику лицензирования лекарственных средств, ужесточив требования к лицензируемым препаратам. — *Прим. ред.*

публикации, украшенные статистическим «макияжем» и статусными регалиями авторов (кандидат медицинских наук, доктор медицинских наук, профессор, академик, ректор вуза, директор НИИ и т.п.), практически удаляют критическую оценку декларируемых выводов читателями (как медиками-исследователями, так и медиками-практиками).

Во второй половине XX в. использование новых лекарственных препаратов и медицинских технологий в зарубежной медицине привело к значительному росту объемов информационных ресурсов. Именно развитие процедур обобщения и концентрации этих ресурсов, в том числе с помощью медицинской статистики, и привело в итоге к появлению доказательной медицины (evidence-based medicine). В отличие от специалистов СССР и России зарубежные медики сделали очень многое по использованию статистики как стандартного инструмента медицинской науки [224]. В 1938 г. была создана Биометрическая секция Американской статистической ассоциации. Затем в 1947 г. в Вудс-Холе (США) проведена Первая международная биометрическая конференция, на которой и было организовано Международное биометрическое общество. Конференции Международного биометрического общества проходили в 1949, 1953, 1958, 1963, 1967 гг. и т.д. В 1978 г. было организовано Международное общество **клинической биостатистики** (ISCB), национальные отделения которого есть в нескольких десятках стран, включая США, Англию, Францию, Италию, Канаду, Испанию, Польшу, Венгрию, Южную Африку, Кению и т.д. А с 1982 г. издательством John Wiley & Sons издается специализированный журнал «**Statistics in Medicine**». В 1998 г. это издательство выпустило 6-томную «**Энциклопедию биостатистики**», содержащую более 2 тыс. статей. При этом за рубежом издают тысячи книг по использованию статистики в разных иных науках. Одним из наиболее популярных таких изданий как раз и является данная книга «Наглядная медицинская статистика». В США издаются статистические журналы «International Journal of Statistics in Medical Research», «Journal of Medical Statistics and Informatics» и т.д.

Наряду с этим в зарубежных университетах имеются многочисленные факультеты подготовки специалистов по теории вероятностей и статистике. Такие специалисты активно выполняют продуктивные статистические анализы не только для медиков, но и для биологов, генетиков, психологов, социологов, экономистов, химиков, технологов. И вот в этой ситуации как раз и возникает вопрос, а почему же эти медики, биологи, генетики и т.д. не сами выполняют продуктивные статистические анализы своих баз данных, а обращаются в проводимых исследованиях за помощью именно к профессионалам по статистическому анализу? Ниже и привожу читателям подробное объяснение этого отношения.

Во-первых, эти медики являются профессионалами высокого уровня, и поэтому они весьма загружены своими технологиями. И они осознают высокую сложность используемой ими медицинской технологии, которую им как раз и необходимо постоянно и существенно улучшать.

Во-вторых, эти профессионалы в области медицины понимают, что в их медицинских технологиях практически все признаки состояния пациентов и признаки проводимых их технологий могут иметь много разных взаимозависимостей. То есть признаки могут иметь парные и многомерные взаимосвязи разной интенсивности. И эти связи могут быть не только линейными, но и других типов.

В-третьих, эти профессионалы в области медицины понимают, что для получения их технологиям продуктивных результатов статистического анализа им как раз и необходимо создавать достаточно объемные базы данных. То есть в этих базах данных нужно использовать большое количество наблюдений (пациентов) и большое количество признаков. И они должны быть и количественные, и качественные, группирующие признаки.

В-четвертых, эти профессионалы в области медицины, самостоятельно применяя самые простые методы статистического анализа, осознают, что столь простые результаты как раз и не могут существенно улучшить их сложные медицинские технологии.

В-пятых, эти профессионалы в области медицины осознают, что, не будучи профессионалами в области математической статистики и пытаясь самостоятельно использовать сложные методы статистического анализа, они получают ошибочные результаты, используя которые в своих медицинских технологиях, могут отрицательно повлиять на свои технологии.

Именно такие профессионалы, осознающие пять аспектов и желающие улучшить свои технологии, привлекают в свои исследования профессионалов по статистическому анализу. Прочие же публикуют примитивные, а часто и ошибочные результаты статистического анализа, отображая этим и свой уровень профессионализма [223]. Одной из причин таких публикаций является отсутствие знаний о возможностях иных, гораздо более продуктивных методов, например, многомерных методов статистического анализа.

Поэтому для понимания целей современных методов статистического анализа рекомендую читателям этой книги внимательно ознакомиться с описаниями этих методов. Однако, поскольку невозможно лишь из одной книги узнать цели практически всех методов статистического анализа, мы и привели в разделе «Приложение к русскому изданию. Библиография от научного редактора» список 455 книжных изданий и статей, из которых 409 — русскоязычные. Многие источники содержат ссылки на интернет-ресурсы. В частности, на сайте Биометрика приведено немало результатов статистического анализа разного уровня.

В.П. Леонов,
редактор сайта Биометрика (http://www.biometrica-tomsk.ru)

Предисловие к четвертому изданию на английском языке

«Наглядная медицинская статистика» (Medical Statistics at a Glance) предназначена для медиков-студентов, медицинских исследователей, аспирантов биомедицинских дисциплин и персонала фармацевтической промышленности. Все эти лица в своей профессиональной практике неизбежно столкнутся с количественными результатами (собственными или чужими), которые потребуют критической оценки и интерпретации, а некоторые из них, конечно, будут вынуждены пройти экзамен по этой наводящей ужас статистике! Надлежащее понимание статистических концепций и методологии неоценимо для этих целей. Будучи прагматиками, мы хотели бы зажечь читателя энтузиазмом в области статистики. Наша цель — дать студенту и исследователю, как и клиницисту, которые сталкиваются со статистическими концепциями в медицинской литературе, книгу, которая логична, легко читается, исчерпывающа и имеет практическое применение.

Мы полагаем, что *«Наглядная медицинская статистика»* будет особенно полезна в качестве дополнительного материала к лекциям по статистике и как руководство с соответствующими ссылками. Структура настоящего, четвертого издания такая же, как и у первых трех. Вместе с другими книгами серии «At a Glance» мы ведем читателя через ряд самостоятельных двух- и трехстраничных тем, причем каждая из них охватывает отдельный аспект медицинской статистики. Из собственного опыта обучения мы узнали и приняли во внимание трудности, с которыми столкнулись наши студенты при изучении медицинской статистики. Имеется разветвленная сеть перекрестных ссылок во всем тексте издания, чтобы помочь читателю увидеть связь между разными приемами исследования. По этой причине мы предпочли ограничить теоретическое содержание книги до уровня, который достаточен для понимания включенных процедур, но все-таки не затмевает практических аспектов их выполнения.

Медицинская статистика — предмет, имеющий широкий диапазон и включающий большое число глав. Мы дали элементарное введение в основополагающие концепции медицинской статистики и руководство по наиболее часто применяемым статистическим процедурам. Эпидемиология тесно связана с медицинской статистикой, поэтому обсуждаются некоторые основные вопросы, относящиеся к разработке исследования и его интерпретации. Включены также темы, которые читатель может найти полезными только изредка, но которые, тем не менее, фундаментальны во многих областях медицинских исследований, например доказательная медицина, систематические обзоры и метаанализ, анализ временных рядов, анализ выживаемости и байесовские методы. Мы объяснили принципы, лежащие в основе этих тем, так, чтобы читатель смог понять и интерпретировать результаты, когда они представлены в литературе.

Основная часть статистических таблиц содержится в приложении A. Neave H.R. (1995) Elementary Statistical Tables, Routledge: London, and Diem K., Lenter C. и Seldrup (1981) Geigy Scientific Tables, 8th rev. and enl. edition, Basle: Ciba-Geigy среди прочих изданий предлагают полную версию, если читателю требуются более точные данные для ручных вычислений.

В четвертом издании появилось новое приложение D, которое содержит руководящие принципы контролируемых рандомизированных испытаний (вопросник и блок-схема CONSORT) и исследований по данным наблюдений. Вопросники CONSORT и STROBE сгенерированы сетью EQUATOR, созданной с целью обеспечения ресурсами и обучения формированию отчетов о результатах исследований в области здравоохранения. Руководящие принципы для представления результатов исследования сейчас доступны для многих других его типов, поэтому мы приводим адреса сайтов в таблице (приложение D) для некоторых проектов такого рода. Приложение D также содержит шаблоны, которые, как мы надеемся, окажутся полезными для количественной и качественной критической оценки доказательств при рандомизированных контрольных испытаниях и исследованиях по данным наблюдений.

Названия глав настоящего четвертого издания совпадают с таковыми в третьем издании. Содержание некоторых из первых 46 глав осталось неизменным в данном новом издании; в другие внесены незначительные изменения с учетом последних достижений, перекрестных ссылок и реструктуризации нового материала. В частности, везде, где возможно, мы даем ссылки на значимые основные инструкции системы EQUATOR.

Как и в третьем издании, мы приводим список целей изучения для каждой главы. Перечень предполагает определенные рамки для оценки понимания материала и успехов в обучении. Если вы сможете выполнить все задачи, помеченные в текущей главе, на удовлетворительном уровне, вы овладеете понятиями, рассмотренными в ней.

Большая часть статистических техник, описанных в книге, сопровождается примерами их практического применения. Мы заменили многие старые примеры из прежних изданий теми, которые соответствуют данным текущих клинических исследований. Получили такие данные для примеров на основании совместных исследований: вместе с нами в них участвовали наши коллеги. В некоторых случаях мы использовали реальные документированные данные из открытых источников. Везде, где это представляется возможным, мы пользуемся тем же самым набором данных в нескольких главах, чтобы отразить реалистичность анализа данных, который

редко ограничивается одной техникой или подходом. Тем не менее, верим, что следует привести формулы, а логику выбранного подхода необходимо объяснить для лучшего понимания. Но при этом опущены детали сложных вычислений, так как большинство наших читателей пользуется компьютерами и маловероятно, чтобы они выполняли вручную даже простые вычисления.

Полагаем, что для читателей особенно важно уметь интерпретировать выходные данные работы компьютерных программных пакетов. Следовательно, там, где это применимо, показаны результаты, извлеченные из выходных компьютерных данных. В некоторых примерах, где, по нашему мнению, возможны затруднения при интерпретации данных, включены (см. приложение С) полные компьютерные распечатки анализа набора данных. Существует много широко используемых статистических наборов. В связи с этим для того, чтобы дать читателю указания о том, каким образом можно изменить выходные данные, мы не ограничились процедурой вывода в рамках какого-либо определенного пакета, а использовали вместо этого четыре хорошо известных — SAS, SPSS, Stata и R.

Известно, что одна из наибольших трудностей, с которой сталкиваются не-статистики, — это выбор подходящего метода статистического анализа. В связи с этим мы привели две древовидные диаграммы, которые помогут выбрать, какой конкретно метод использовать в данной ситуации, и легко находить в книге необходимую технологию. Эти диаграммы изображены на внутренней стороне обложки книги.

Читатель может оценить свои успехи в самообучении, выполняя интерактивные упражнения на нашем сайте http://www.medstatsaag.com, а также тесты с множественным выбором и отвечая на структурированные вопросы с шаблонными ответами при использовании нашего пособия «Наглядная медицинская статистика. Сборник упражнений». Этот сайт также содержит полный набор ссылок (некоторые из них непосредственно связаны с базой данных Medline), чтобы расширить ссылки, указанные в тексте, и обеспечить себя дополнительной полезной информацией для примеров. Для тех читателей, которые желают получить большее понимание в специфических областях медицинской статистики, мы можем рекомендовать следующие книги.

- *Altman D.G.* Practical Statistics for Medical Research. London: Chapman and Hall/CRC, 1991.
- *Armitage P., Berry G., Matthews J.F.N.* Statistical Methods in Medical Research. 4th ed. Oxford: Blackwell Science, 2001.
- *Kirkwood B.R., Sterne J.A.C.* Essential Medical Statistics. 2nd ed. Oxford: Blackwell Publishing, 2003.
- *Pocock S.J.* Clinical Trials: A Practical Approach. Chichester: Wiley, 1983.

Мы чрезвычайно благодарны Марку Гилторпу и Джонатану Стерну, которые сделали неоценимые комментарии и предложения по ряду аспектов второго издания, а также Ричарду Моррису, Фионе Лампе, Шаку Хаджату и Абулу Басару за их обсуждения первого издания. Мы желаем поблагодарить каждого, кто помог нам, обеспечивая данные для примеров. Естественно, мы берем полную ответственность за любые ошибки, которые остаются в тексте или в примерах. Мы также хотели бы поблагодарить Майка, Джеральда, Нину, Эндрю и Карен, терпеливо и хладнокровно переносивших нашу увлеченность первыми тремя изданиями и переживших вместе с нами все испытания и огорчения этого четвертого издания.

Авива Петри,
Кэролайн Сэбин,
Лондон

Сайт партнеров: www.medstatsaag.com

Теоретические распределения: другие распределения 8

Цели изучения

К концу этой главы вы должны овладеть следующими знаниями.

- Перечисление важнейших свойств *t*-распределения Стьюдента, χ^2-распределения Пирсона, *F*-распределения Фишера–Снедекора и логнормального распределения вероятностей.
- Объяснение, когда каждое из этих распределений вероятности наиболее полезно.
- Перечисление важнейших свойств биномиального распределения и распределения Пуассона.
- Объяснение, когда наиболее полезны биномиальное распределение и распределение Пуассона.

Несколько слов для поддержки духа

Не стоит волноваться, если теория, лежащая в основе распределения вероятности, вам покажется сложной. Наш опыт показывает, что единственное, что вы хотели бы знать, — это когда и как применять эти распределения. Поэтому мы изложили основы и опустили уравнения, которые характеризуют распределения вероятности. И вы поймете, что единственное, что вам нужно, — это освоить основные понятия, терминологию и, возможно, знать, как работать с таблицами.

Непрерывное распределение вероятностей

Эти распределения основаны на непрерывных случайных переменных. Часто это не непосредственно измеряемая переменная, которая отвечает такому распределению, а параметр, **статистика**, полученная из этой переменной. Общая площадь под кривой функции плотности распределения вероятности есть сумма вероятностей всех возможных значений, и она равна 1 (глава 7). Мы рассмотрели нормальное распределение в главе 7; в этой главе мы опишем другие наиболее известные распределения.

t-распределение (приложение A2, рис. 8.1)

- Получено Вильямом Госсетом, который публиковался под псевдонимом Student (Студент)[1], поэтому его часто называют ***t*-распределением Стьюдента**.

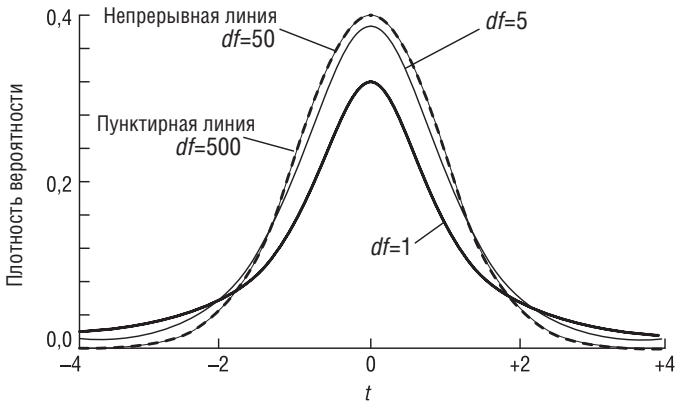

Рис. 8.1. *t*-распределение со степенями свободы (*df*)=1; 5; 50 и 500

[1] Статья была опубликованна в 1908 г. в журнале Biometrika (см. http://www.biometrica-tomsk.ru/student_1908_1.pdf и http://www.biometrica-tomsk.ru/student_1908_2.pdf). — *Прим. ред.*

- Параметрами, которые характеризуют *t*-распределение, являются **степени свободы** (*df*), так, мы сможем начертить функцию плотности распределения вероятности только в том случае, если мы будем знать уравнение *t*-распределения и степени свободы. В главе 11 мы рассмотрим степени свободы; обратите внимание, что они часто выражаются через объем выборки.
- Форма подобна такой же для стандартизованного нормального распределения, но более приплюснута и с более длинными хвостами. Форма приближается к нормальной кривой, по мере того как увеличиваются степени свободы.
- В частности, его применяют для вычисления доверительных интервалов и исследования гипотез с одной или двумя средними (главы 19–21).

Хи-квадрат (χ^2) распределения Пирсона (приложение A3, рис. 8.2)

- Является скошенным вправо распределением, принимающим только положительные значения.
- Характеризуется **степенями свободы** (глава 11).
- Форма зависит от числа степеней свободы; становится более симметричным и приближается к нормальному с их ростом.
- Особенно часто используется для анализа категориальных данных (главы 23–25).

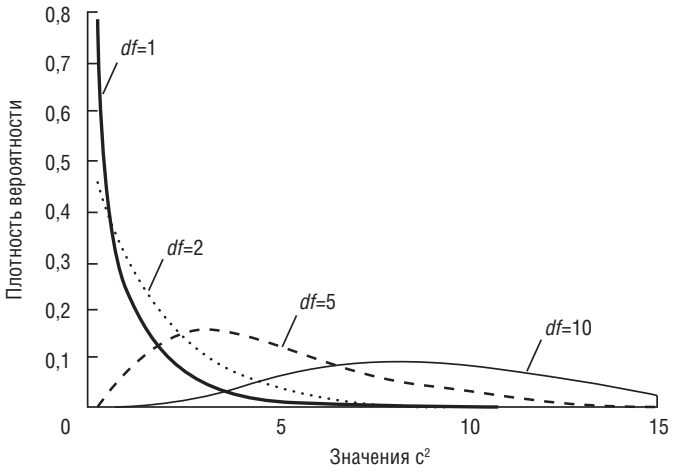

Рис. 8.2. χ^2 распределения Пирсона со степенями свободы (*df*)=1, 2, 5 и 10

F-распределение (приложение A5)

- Является скошенным вправо.
- Определяется как отношение. Распределение отношения двух оценок дисперсий, вычисленных для нормально распределенных данных, аппроксимируются *F*-распределением.
- Два параметра, которые характеризуют его, — степени свободы (глава 11) числителя и знаменателя отношения.
- *F*-распределение особенно полезно для сравнения двух дисперсий (глава 35) и более чем двух средних при использовании дисперсионного анализа (ANOVA) (глава 22).

Логнормальное распределение

- Это распределение вероятности случайной переменной, логарифм которого (по основанию 10 или *e* — основание натурального логарифма) имеет нормальное распределение.
- Сильно скошено вправо (рис. 8.3, а).
- Если мы возьмем логарифмы наших исходных данных, которые скошены вправо, мы создадим эмпирическое распределение,

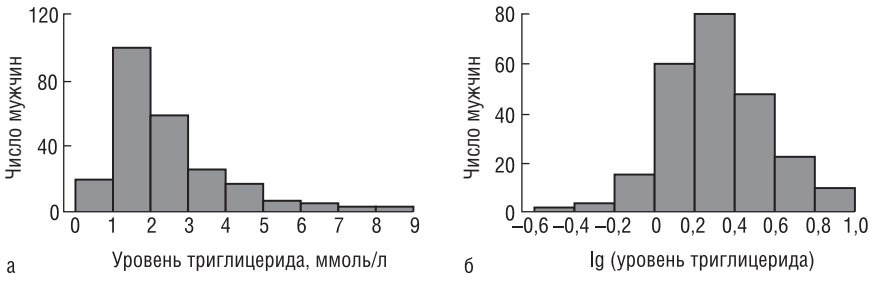

Рис. 8.3. Логнормальное распределение уровня триглицерида у 232 мужчин с заболеваниями сердца (глава 19) (а). Почти нормальное распределение $\log_{10}$ (уровень триглицерида) (б)

которое почти нормальное (рис. 8.3, б), и тогда наши данные соответствуют приближенно логнормальному распределению.

- Многие переменные в медицине имеют логнормальное распределение. Мы можем использовать свойства нормального распределения (глава 7), для того чтобы сделать выводы этих переменных после логарифмического преобразования данных.
- Если набор данных имеет логнормальное распределение, мы используем среднее геометрическое (глава 5) как обобщающий показатель положения.

Дискретные распределения вероятностей

Случайная переменная, которая имеет такое распределение вероятности, является дискретной. Сумма вероятностей всех возможных взаимоисключающих событий равна 1.

Биноминальное распределение

- Предположим, в данной ситуации существует только два результата — «успех» и «неудача». Например, нас интересует, забеременеет или нет женщина при экстракорпоральном оплодотворении (IVF). Если мы посмотрим на n=100, не имеющих отношение друг к другу женщин, подвергающихся IVF (каждая с той же вероятностью беременности), то биноминальная случайная переменная — это наблюдаемое количество зачатий. Часто это понятие объясняется в терминах n независимых повторных испытаний (например, 100 раз подбросить монету), при которых результатом будет являться либо успех (например, орел), либо неудача.
- Два параметра, которые описывают биноминальное распределение, — это n — количество индивидуумов в выборке (или повторения испытания) и p — точная вероятность успеха для каждого индивидуума (или при каждом испытании).
- Среднее (значение для случайной переменной, которое мы ожидаем, если мы осматриваем n индивидуумов или повторим испытание n раз) — это np. Дисперсия — $np(1-p)$.
- Когда n невелико, распределение будет скошено вправо, если $p<0,5$, и влево, если $p>0,5$. Распределение становится более симметричным, по мере того как объем выборки будет увеличиваться (рис. 8.4), и приблизится к нормальному распределению в том случае, если np и $n(1-p)$—>5.
- Мы можем использовать свойства биномиального распределения, для того чтобы сделать выводы относительно **пропорций**. Особенно часто мы используем аппроксимацию биномиального распределения с помощью нормального распределения при анализе пропорций (долей).

Распределение Пуассона

- Пуассоновская случайная переменная — это **число** событий, которые происходят независимо и случайно во времени или пространстве со средней интенсивностью **μ**. Например, количество госпитализаций в день типично отвечает распределению Пуассона. Мы используем распределения Пуассона, для того чтобы вычислить вероятность конкретного количества госпитализаций в любой отдельный день.
- Параметр, которым описывают распределение Пуассона, — это **среднее**, то есть средняя интенсивность, **μ**.
- **Среднее** равняется **дисперсии** в распределении Пуассона.
- Если среднее маленькое, то распределение будет скошено вправо и будет становиться более симметричным; по мере того как среднее будет увеличиваться, оно приближается к нормальному распределению.

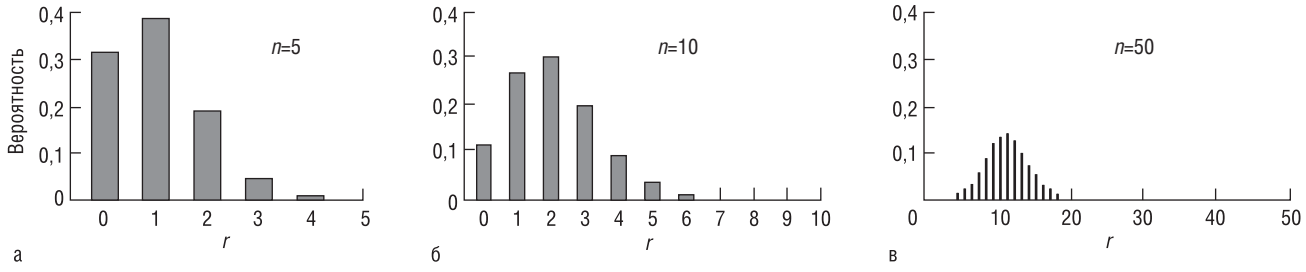

Рис. 8.4. Биноминальное распределение показывает количество успехов r, когда вероятность успеха p=0,20 для объема выборки n=5 (а), n=10 (б) и n=50 (в) (NB! В главе 23 наблюдаемая серораспространенность HHV-8 была p=0,185≈0,2 и объем выборки был 271: для пропорции было использовано приближение нормальным распределением)

Преобразования 9

Цели изучения

К концу этой главы вы должны овладеть следующими знаниями.

- Описание ситуаций, в которых преобразование данных может быть полезным.
- Объяснение, как преобразовать набор данных.
- Объяснение, когда применять и что достигается логарифмированием, извлечением квадратного корня, получением обратной величины, возведением в квадрат и логит-преобразованием.
- Описание, как интерпретировать итоговые результаты, полученные с помощью логарифмирования данных, после их обращения к оригинальной шкале.

Зачем преобразовывать?

В нашем исследовании наблюдения могут не подчиняться требованиям предполагаемого статистического анализа (глава 35).

- Переменная может иметь ненормальное распределение — требование, обязательное для множества различных параметрических методов анализа (например, *t*-критерий Стьюдента, ANOVA и т.д.).
- Рассеяния (размах) признака в наблюдениях каждой из групп ряда могут быть разными, но равенство дисперсий — это необходимое условие корректного использования **параметра** при сравнении средних с помощью *t*-критерия Стьюдента и классического дисперсионного анализа (ANOVA) (главы 21–22).
- Две переменные не могут быть линейно зависимы (**линейность** — это требование во многих видах регрессионного анализа, см. главы 27–33 и 42).

Очень часто полезно **преобразовывать** данные, для того чтобы удовлетворить требованиям, которые лежат в основе предлагаемых статистических методов[1].

Как мы преобразовываем?

Мы превращаем исходные данные в преобразованные, используя одно и то же математическое преобразование для каждого наблюдения. Предположим, что у нас есть n наблюдения ($y_1, y_2, ..., y_n$) с переменной y и мы принимаем решение, что нам подходит логарифмическое преобразование. Мы берем логарифм каждого наблюдения, для того чтобы образовать ($\log y_1, \log y_2, ..., \log y_n$). Если мы назовем преобразованную переменную z, тогда $z = \log y_i$ для каждой i ($i = 1, 2, ..., n$) и наше преобразование данных можно записать ($z_1, z_2, ... z_n$).

Мы проверяем, достигло ли это преобразование своей цели при создании набора данных, который удовлетворяет предположениям запланированного статистического анализа, и переходим к анализу преобразования данных ($z_1, z_2, ..., z_n$). Мы часто делаем обратные преобразования обобщающих мер (таких, как среднее) до первоначальной размерности; те выводы, которые мы сделали из гипотез по нашим тестам (глава 17) на преобразованных данных, применимы для исходных данных[2].

[1] При этом важно помнить, что результат применения статистического критерия к преобразованной переменной нельзя априорно переносить на исходную переменную. Далее преобразование может изменить и размерность числовой переменной, что затрудняет смысловую интерпретацию новой переменной. Например, систолическое давление имеет размерность (мм рт.ст.). После возведения в квадрат этой переменной размерность станет (мм рт.ст.)^2. Каков физический смысл это новой переменной? — *Прим. ред.*

[2] Такое согласие наблюдается далеко не всегда, и потому оно требует дополнительной проверки. — *Прим. ред.*

Типичные преобразования

Логарифмическое преобразование, $z = \log(y)$

При логарифмическом преобразовании данных мы можем выбирать, взять логарифмы по основанию 10 ($\log_{10}(y)$ — десятичный логарифм) или по основанию e ($\log_e(y) = 1n(y)$ — натуральный или неперовский логарифм), но они должны быть одинаковы для отдельной переменной в наборе данных. Обратите внимание, что мы не можем брать логарифм отрицательного числа или нуля. Обратное преобразование логарифма (потенцирование) называется антилогарифмом; антилогарифм неперовского логарифма есть экспонента — e.

- Если распределение y скошено вправо, преобразование $z = \log(y)$ часто дает в результате приближенно **нормальное распределение** (рис. 9.1, а). Тогда y имеет логнормальное распределение (глава 8).
- Если существует экспоненциальное соотношение между y и другой переменной x, такое, что конец кривой загибается вверх, в то время как y (по вертикальной оси) наносится соответственно x (по горизонтальной оси), то соотношение между $z = \log(y)$ и x приблизительно **линейное** (рис. 9.1, б).
- Предположим, что у нас есть различные группы наблюдений, причем включая измерения непрерывной переменной y. Мы можем обнаружить, что группы, которые имеют более высокие значения y, имеют и большие вариации. В частности, если коэффициент вариации (стандартное отклонение, деленное на среднее), постоянен для всех групп, преобразование логарифма $z = \log(y)$ создает группы с **равными дисперсиями** (рис. 9.1, в).

В медицине часто применяется логарифмическое преобразование из-за логической интерпретации и потому, что многие переменные имеют скошенное вправо распределение. Например, если исходные данные подвергнуть логарифмическому преобразованию, тогда разница между двумя средними величинами на логарифмической шкале равна отношению двух соответствующих средних в исходной шкале. Антилогарифмы границ 95% доверительного интервала (глава 11) для среднего значения логарифмически преобразованных данных дают границы 95% доверительного интервала для геометрического среднего. Если мы используем для независимых переменных (предикторов) в регрессионном анализе (глава 29) логарифмирование по основанию 10, увеличение переменной в логарифмической шкале на единицу представляет собой 10-кратное увеличение переменной в исходной, оригинальной шкале. Обратите внимание, что логарифмическое преобразование в регрессии зависимой, результирующей переменной дает возможность обратного преобразования регрессионных коэффициентов и мультипликативного эффекта, а не суммирования эффекта в оригинальной, исходной шкале (см. главы 30 и 31).

Преобразование квадратного корня $z = \sqrt{y}$

Это преобразование имеет свойства такие же, как и логпреобразование, хотя результаты, после того как они были преобразованы обратно, объяснить сложнее. В дополнение к способностям **нормализации** и **линеаризации**, эффективно использовать преобразование, **стабилизирующее дисперсию**, если дисперсия возрастает при увеличении значений y, то есть тогда дисперсия, деленная на среднее, постоянна. Мы применяем преобразование квадратного корня, если y — это количество редких явлений, встречающихся во времени и пространстве, то есть это пуассоновская переменная (глава 8). Помните, что мы не можем извлечь квадратный корень из отрицательного числа.

Обратное преобразование $z = 1/y$

Мы часто применяем обратное преобразование к периодам жизни (выживаемости), если не используем специальные методы для анализа выживаемости (глава 44). Обратное преобразование имеет

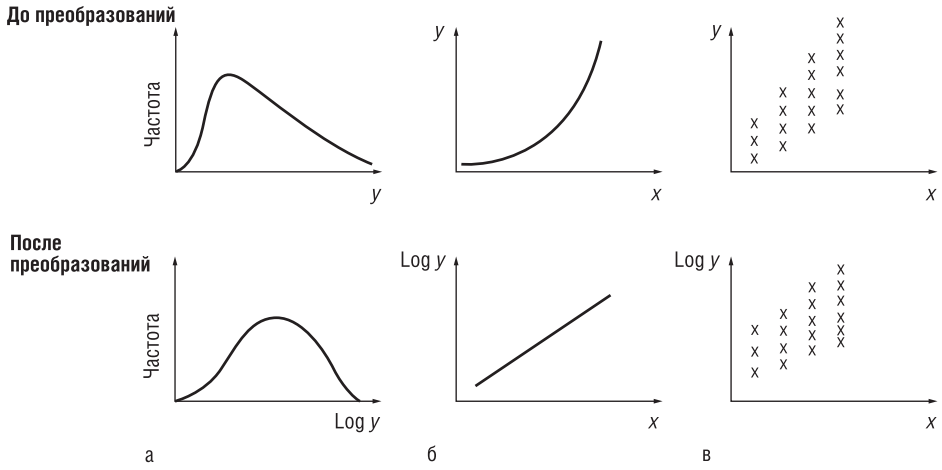

Рис. 9.1. Свойства логарифмического преобразования: нормализация (а), линеаризация (б), выравнивание дисперсий (в)

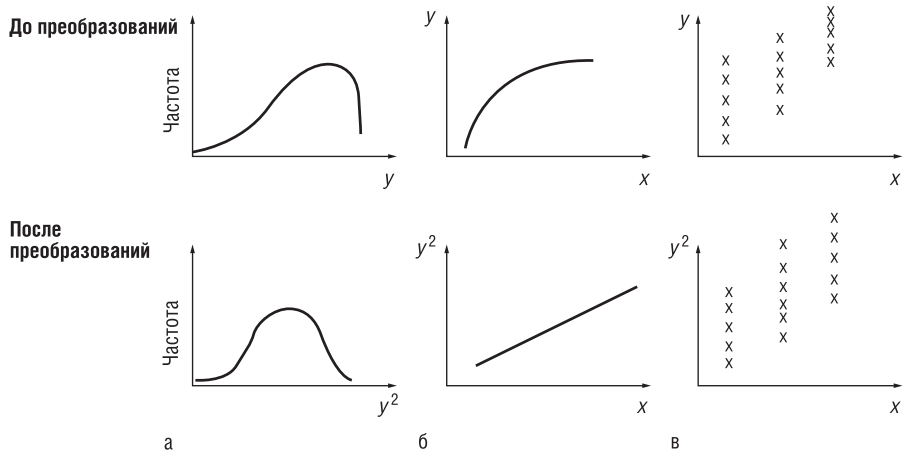

Рис. 9.2. Эффект квадратичного преобразования: нормализация (а), линеаризация (б), выравнивание дисперсий (в)

такие же свойства, как и лог-преобразование. В дополнение к способностям **нормализации** и **линеаризации**, оно гораздо эффективнее для **стабилизации дисперсии**, чем логарифмическое преобразование, если дисперсия очень заметно увеличивается при увеличении значений y, то есть для стабилизации частного от деления дисперсии на среднее. Заметьте, что мы не можем делить на нуль.

Квадратичное преобразование $z = y^2$

Квадратичное преобразование достигает результат, обратный лог-преобразованию.

- Если y скошено влево, то распределение $z = y^2$ часто является приближенно **нормальным** (рис. 9.2, а).
- Если соотношение между двумя переменными x и y такое, что, когда мы наносим y против x, кривая загибается вниз, тогда соотношение между $z = y^2$ и x почти **линейное** (рис. 9.2, б).
- Если дисперсия непрерывной переменной y уменьшается, по мере того как значение y увеличивается, тогда квадратное преобразование $z = y^2$ **стабилизирует дисперсию** (рис. 9.2, в).

Логит (логистическое) преобразование $z = \ln \frac{p}{1-p}$

Это преобразование, которое мы наиболее часто применяем к каждой пропорции, доле p, в наборе пропорций. Мы не можем применять логистическое преобразование в том случае, если p=0 либо p=1, потому что соответствующие значения логита — это $-\infty$ и $+\infty$. Единственное решение — это взять p как $1/(2n)$ вместо 0 и как $\{1-1/(2n)\}$ вместо 1. Оно линеаризует сигмовидную кривую (рис. 9.3), см. главу 30 по использованию логит-преобразования в логистической регрессии[1].

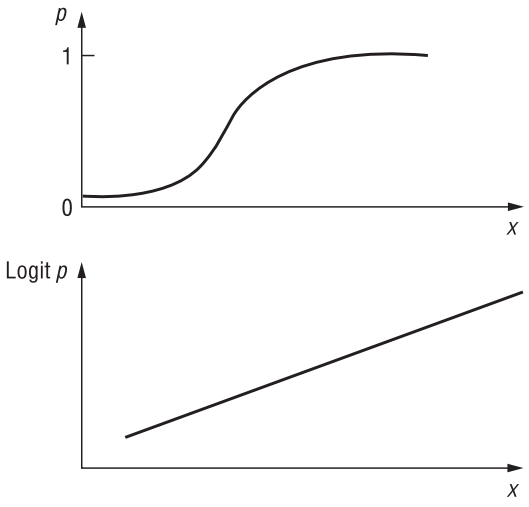

Рис. 9.3. Эффект логистического преобразования сигмовидной кривой

[1] Достаточно подробное описание специфики использования этого мощного метода с большим количеством реальных примеров приведено в статье «Логистическая регрессия в медицине и биологии» (см. http://www.biometrica-tomsk.ru/logit_0.htm).