Лапач С.Н., Пасечник М.Ф., Чубенко А.В.

Статистические методы в фармакологии и маркетинге фармацевтического рынка

Монография

Киев — 1999

ББК 22.18
Л24
УДК 519.21:614.27

Утверждено к печати Ученым Советом Национального технического университета Украины "Киевский политехнический институт", Ученым Советом института "Киевский бизнес-колледж".

Рецензенты:

Коваленко А.С. доктор медицинских наук, заведующий отделом Международного учебно-научного центра информационных технологий и систем НАНУ и Министерства образования.

Воронин А.Н. доктор технических наук, ведущий научный сотрудник Института космических исследований НАНУ и НКАУ.

Лапач С.Н., Пасічник М.Ф., Чубенко А.В.

Л24 Статистичні методи в фармакології і маркетингу фармацевтичного ринку.– К.: ЗАТ "Укрспецмонтажпроект", 1999.– 312 с.

Розглянуті статистичні методи, які використовуються при розв`язанні типових задач в фармакології та маркетингу. Досліджуються питання побудови математичних моделей по експериментальним даним (лінійна по параметрам регресія). Викладена розроблена технологія побудови моделей, що дозволяє автоматизувати отримання структури рівняння регресії: генерація робастних планів експерименту, алгоритми побудови структури рівняння регресії. Приведений детальний опис розв`язання ряду задач з використанням статистичних методів в маркетингу та доклінічних дослідженнях.

Книга буде корисна всім, хто займається розробкою та випробуванням лікарських препаратів, а також аспірантам та студентам.

Лапач С.Н., Пасечник М.Ф., Чубенко А.В.

Л24 Статистические методы в фармакологии и маркетинге фармацевтического рынка.– К.: ЗАТ "Укрспецмонтажпроект", 1999.– 312 с.

Рассмотрены статистические методы, применяемые для решения типовых задач в фармакологии и маркетинге. Исследуются вопросы построения математических моделей по экспериментальным данным (линейная по параметрам регрессия). Изложена разработанная технология построения моделей, позволяющая автоматизировать получение структуры уравнения регрессии: генерация робастных планов эксперимента, алгоритмы получения структуры уравнения регрессии. Приводится подробное описание решение ряда задач с применением статистических методов и многокритериальной оптимизации в маркетинге и доклинических исследованиях.

Книга будет полезна всем кто применяет статистические методы в практической деятельности при создании и испытаниях препаратов, а также аспирантам и студентам.

Научное издание

ISBN 966-74-12-09-1

ББК 22.18

© Лапач С.Н., Пасічник М.Ф., Чубенко А.В.

© ЗАО "Укрспецмонтажпроект"

Наукове видання

Сергій Миколайович Лапач

Михайло Францович Пасічник

Анатолій Васильович Чубенко

Статистичні методи в фармакології і маркетингу фармацевтичного ринку

Монографія

(На російській мові)

Підписано до друку 21.04.99. Формат 60x80 $^{1}/_{16}$. Папір друкарський. Друк офсетний. Умовно друкованих аркушів 14,02. Замовлення 17с.

Тираж 500 прим. Ціна договірна.

ЗАТ "Укрспецмонтажпроект"

Сведения об авторах

Лапач Сергей Николаевич
Научный сотрудник Фармакологического Комитета, преподаватель Национального технического университета “Киевский политехнический институт”. Автор большого количества работ (в том числе одной монографии) в области разработки и применения экспериментально-статистических методов. Разработчик программных средств ПРИАМ (планирование, регрессия и анализ моделей) и OptimeChoice (многокритериальная оптимизация). Основные научные интересы — разработка и применение методов многомерного статистического анализа и многокритериальной оптимизации.

Пасичник Михаил Францевич
Президент фармацевтической компании «Фалби». Окончил Киевский политехнический институт и Украинскую фармацевтическую академию. Автор многих работ по фармакологии, фармации, управлению и маркетингу фармацевтического бизнеса в Украине. В настоящее время основная сфера научных интересов — разработка современных информационных технологий с целью интенсификации внедрения лекарственных средств.

Чубенко Анатолий Васильевич
Ведущий научный сотрудник Фармакологического комитета, к.м.н., заведующий отделом Института фармакологии и токсикологии. Автор большого количества работ (в том числе одной монографии) в области фармакологии и токсикологии. Область научных интересов — разработка методов количественной оценки фармакологического эффекта, современные информационные технологии фармацевтического рынка.

СОДЕРЖАНИЕ

Аннотация

Современная медицина и фармакология немыслимы без использования статистических методов. Их важность увеличивается в связи с принятием требований GMP, GLP и GCP в качестве определенных международных стандартов, используемых во всем жизненном цикле лекарственных средств от разработки до прекращения применения в медицинской практике.

Предлагаемая книга имеет следующую структуру: первый и второй разделы содержат краткое описание основных задач, проблем и методов статистического анализа. Остальные разделы посвящены построению и использованию линейных по параметрам и в общем случае нелинейных по факторам регрессионных моделей. В последнем разделе приведены примеры решения типовых задач с применением статистических методов и многокритериальной оптимизации в маркетинге и доклинических исследованиях;.

Линейный относительно параметров регрессионный анализ является одной из самых широко применяемых статистических процедур. Вместе с тем, существует ряд причин, препятствующих его эффективному использованию и правильной интерпретации результатов. Они связаны с тем, что с ростом сложности и размерности решаемых задач проблемы невыполнения предпосылок и допущений регрессионного анализа стали играть большую роль, приводя в ряде случаев к невозможности решения или неправильному решению задачи. К наиболее важным предпосылкам относятся необходимость знать структуру модели до генерации плана и расчета коэффициентов и проблема определения частной структуры уравнения регрессии (часто называемой проблемой выбора наилучшего подмножества регрессоров).

В данной работе рассматриваются вопросы построения математических моделей по экспериментальным данным (линейная по параметрам регрессия). Изложена разработанная технология построения моделей, позволяющая автоматизировать получение структуры уравнения регрессии. Исследованы вопросы построения робастных планов эксперимента и алгоритмов получения структуры уравнения регрессии. В связи с тем, что работа ориентирована на специалистов-прикладников, ей придан такой вид, чтобы ее результатами можно было воспользоваться в практической деятельности. Известно, что ряд работ, имеющий важное значение для прикладников, практически был не замечен большинством специалистов (например, Бродский В.З. «Многофакторные регулярные планы», Демиденко Е.З. «Линейная и нелинейная регрессия») в связи с высоким уровнем абстракции и насыщенностью собственно математическими проблемами. Поэтому в данной работе основное внимание уделялось полученным научным и прикладным результатам и выводам, а также их интерпретации. Для ряда методов и алгоритмов не

приведено подробного описания в связи с тем, что их использования без соответствующего программного обеспечения, например, рарзработанного программного средства “Планирование, регрессия и анализ моделей” (ПС ПРИАМ), нереально. Описан весь цикл получения регрессионных моделей, включая смежные вопросы (кластерный анализ, вычислительный эксперимент, оптимизация, в том числе многокритериальная), а также введение в статистические методы (случайные величины, проверка гипотез, корреляционный анализ и пр.), что позволяет пользоваться книгой как пособием при изучении статистических методов и при решении задач.

ВВЕДЕНИЕ

С глубокой древности врачи пытались на основании наблюдений делать выводы о влиянии различных факторов на здоровье человека, причины и течение заболеваний. Со времени бурного развития статистики (вторая четверть XIX века), когда Адольф Кетле назвал ее социальной физикой, медицина привлекла ее аппарат для решения своих задач. Применение статистических методов в биологии и медицине приобрело за прошедшее с того времени более чем столетие такой размах, что было выделено в специальную научную дисциплину, называемую ***биометрикой***.

Особенностью современного этапа развития современного естествознания является математизация, неотъемлемой частью которой выступает использование статистических методов для проверки выдвинутых гипотез, обоснованного формирования выборок, построения математических моделей различных явлений и процессов. Практически нет такого класса методов статистического анализа, который не нашел бы применения в медицине.

Однако результаты применения статистических методов не оправдали надежд энтузиастов прошлого века. Сложность использования этих методов в медицине связана с целым рядом объективных причин:

- сильная изменчивость исследуемых признаков ввиду влияния очень большого количества неуправляемых и неконтролируемых факторов;
- проблемы в формировании выборок (или планов экспериментов) требуемого объема и структуры;
- влияние психологических установок и воздействий на результаты испытаний;
- измерение многих важных показателей в неколичественных шкалах — обычно это шкалы классификации и порядка.
- трудности в освоении методов медицинскими работниками, т.к. для решения задач требуются самые современные и мощные средства планирования экспериментов и проведения вычислений.

Несмотря на вышеуказанные сложности, статистические методы заняли прочное место в арсенале современной медицины и фармакологии.

Так, использование планирования эксперимента и регрессионного анализа при проведении доклинических исследований позволяет получить многофакторные зависимости эффекта от дозы, состава препарата, времени применения и необходимых других факторов, а затем использовать их для многокритериальной оптимизации состава и схемы применения препарата при заданных ограничениях на затрачиваемые ресурсы.

В клинике статистические методы служат важным вспомогательным средством получения информации: сравнение эффективности различных видов терапии, влияние различных антропометрических, социальных и

экологических факторов на распространенность и течение заболеваний и др. Одной из важных есть проблема: являются ли определенные побочные эффекты (и даже смерть) следствием применения конкретного препарата в тех случаях, когда это возможно только путем анализа статистических данных. Большое влияние статистические методы оказывают в клинике на формы организации выборок, по которым производится анализ. Это связано с тем, что выборки, полученные использованием фактической совокупности больных в определенной больнице или клинике, в общем случае не является случайными, репрезентативными и однородными. Причины в том, что различные факторы, влияющие на отбор пациентов, приводят к формированию определенной группы больных, которые не являются репрезентативными представителями генеральной совокупности. При проведении клинических исследований наиболее важно обеспечить следующие свойства выборки.

Однородность. В выборке влияние изучаемой совокупности факторов на интересующие признаки не должно противоречить друг другу. То есть при исследовании влияния кофе в этой выборке испытуемых не должно одновременно быть людей, которых кофе возбуждает, и тех, которых кофе клонит в сон[1]. В ряде случаев причины неоднородности могут быть неизвестны и поэтому перед анализом данных желательна проверка выборки методами кластерного анализа.

Структурное соответствие. В тех случаях, когда мы сравниваем некоторые параметры двух выборок, необходимо обеспечить, чтобы в сравниваемых выборках распределение частот влияющих факторов (пол, возраст, серьезность заболевания и пр.) было одинаковым.

Совпадение условий наблюдений. Условия наблюдения для отдельных элементов выборки или для двух сравниваемых выборок должны совпадать. Наилучшим способом обеспечения этого свойства является ***двойной слепой метод***, при котором ни пациент, ни врач, ни средний медицинский персонал не знает какие лекарства или плацебо выдаются конкретному больному. Это позволяет избавится от эффекта внушаемости (влияние которого возможно на 30 — 50 % пациентов) и эффекта предубежденности.

Использование статистических методов в технологии производства лекарственных препаратов позволяет проектировать оптимальные по множеству критериев (технических, экономических, экологических, фармакологических) технологические процессы, контролировать качество

[1] Это достаточно хорошо известно на примере чувствительности людей к фенилтиокабамиду. Одни считают его очень горьким, другие не чувствуют его вкуса вообще при концентрациях в несколько тысяч раз больше, причем это зависит только от особенностей субъекта. При этом больше всего людей нечувствительных к данному химикату среди японцев, китайцев и северо-американских индейцев.

получаемого сырья, выпускаемой продукции, настройки автоматических линий.

Рассмотрение лечебных учреждений как систем массового обслуживания позволяет с использованием статистических методов формировать таким образом структуру и организацию обслуживания населения этими учреждениями, чтобы обеспечивать требуемый качественный и количественный уровень обслуживания при минимальных затратах на его поддержание.

Важное место занимают в маркетинговых исследованиях, позволяющих определить потребности в лекарственных препаратах, тенденции замены одних поколений препаратов другими, наилучшую тактику продвижения препаратов на рынок.

Создание различных баз данных и автоматизированных медицинских систем, накапливающих информацию, остро ставит вопрос о использовании статистических методов для ее обработки. Без этого огромное количество информации становится бесполезным мусором, хранящимся во все возрастающих количествах. Наиболее частыми задачами для таких данных являются распознавание, т.е. разбиение патологических состояний, процесов на классы и определение принадлежности конкретного больного к определенному классу, формирование достаточного набора признаков для диагностики состояния, а также оценка эффективности лечения и определения оптимального лечения в зависимости от особенности патологии и больного.

Следует отметить, что принятие норм GCP, GMP, GLP требует широкого применение статистических методов[2].

В GCP (надлежащая клиническая практика) особое внимание уделяется статистике, особенно подготовке экспериментов, использованию рандомизации и слепого метода при проведении исследований, а также контрольных карт для контроля состояния измерительной аппаратуры и материалов (четвертая и косвенно пятая главы).

GLP (надлежащая производственная практика) особое внимание уделяет контролю качества сырья, конечной продукции и самого технологического процесса, что невозможно без широкого применения статистических методов контроля качества (шестая и частично пятая главы).

Соблюдение принципов GLP (надлежащая лабораторная практика) требует использования статистических методов как для контроля материалов и обородования, так и методически правильной организации эксперимента и формирования научно обоснованных выводов.

Таким образом, статистические методы становятся необходимым инструментом в медицине в целом и в фармакологии в частности. В данной

[2] Flfin Spritn, Therese Dupin-Spriet Good practice of clinical drug trials. KARGER, 1997.— 140 p.; Лицензирование в Европейском союзе: фармацевтический сектор.— К.: Морион-ЛТД, 1998.— 384 с.

книге авторами сделана попытка достичь следующих целей:

- дать читателюобщее представление о спектре статистических методов и задачах его решения;
- представить оригинальную технологию планирования экспериментов и построения регрессионных моделей, которая может быть эффективно использована при проведении доклинических испытаний и анализу комбинированного действия;
- описать разрабатываемую систему поддержки решения маркетинговых задач, базирующуюся на статистических методах и многокритериальной оптимизации.

Есть три источника знания — авторитет, разум и опыт. Однако авторитет недостаточен, если у него нет разумного основания. И разум один не может отличить софизма от настоящего доказательства, если он не может оправдать свои выводы опытом.

Роджер Бэкон

В древнее время задачи ставили боги, как, например, задача удвоения куба — при измерении размеров Дельфийского жертвенника. Потом настал второй период, когда задачи ставили полубоги: Ньютон, Эйлер, Лагранж. Теперь третий период, когда задачи ставит практика.

П.Л. Чебышев

1. Некоторые общие понятия статистики

Эффективность экономики определяется тем, насколько рационально используются ресурсы. Опыт развитых стран показывает, что из всех видов ресурсов наибольший вклад в экономику вносят человеческий интеллект и информация[3].

Традиционно принимается существование двух путей познания окружающего мира с целью его использования в практических целях. Один это накопление эмпирических сведений и их использование без понимания их сути. Второй — это проникновение в сущность явления, теоретическое вскрытие механизмов, явлений и затем осознанное их использование. Ярким представителем первого направления можно назвать кораблестроителя П.А. Титова, второго — инженера В.Г. Шухова и академика А.Н. Крылова. Эти пути продолжают сосуществовать и в наше время. Причем, первый считается уровнем познания ремесленника или рабочего, а второй ученого. В связи с этим они постоянно

[3] Один килограмм стали в США стоит 7 центов (1987 г.), один килограмм массы автомашины — 7 долларов, самолета — 700 долларов, а один килограм интегральных схем — 7000 доларов. То есть знания, «закачанные» в технологию и продукцию, представляют наивысшую коммерческую ценность, дают наибольший экономический эффект (Миронов В. Третий ресурс // Инженер.— 1992.— № 1.— С. 3-4.)

противопоставляются. Такое представление, обусловленное успехами теоретической физики и применением математики в технике в первой трети нашего века, сильно пошатнулось в последние десятилетия.

Хотелось бы обратить внимание, что практически все основные физические законы получены на основании обобщения эмпирического опыта. Они имеют ограниченное использование. Каждый включает обычно не более 2 — 3 независимых переменных и одну зависимую и выполняются лишь приближенно, с точностью до некоторой случайной ошибки, которая зависит от множества различных факторов[4].

В настоящее время технические и технологические объекты и процессы (а также процессы из области биологии, экологии и пр.) значительно отличаются от тех, с которыми приходилось иметь дело еще каких-нибудь 30 лет назад. Это связано с сильным усложнением используемых процессов, явлений; относительно быстрой сменой и внедрением в производство новых процессов, материалов и т.д.

Сложность этих объектов усугубляется невозможностью традиционного (однофакторного) их изучения. Это связано не только с тем, что не хватает ресурсов на такое исследование, при котором все факторы фиксируются, кроме одного, влияние которого на процесс исследуется, но и с тем, что соединить эти однофакторные зависимости в единую картину процесса не удается, т.к. не учитываются взаимодействия факторов. Тут можно провести аналогию с известной притчей о слепцах, которые ощупывали слона (ранее никогда не видев его) за хвост, хобот, ногу, бок и говорили, что слон похож на веревку, трубу, колону и стену. Как по этим описаниям не представить слона, так и по однофакторным исследованиям невозможно создать полноценную картину о процессе или явлении.

Все это привело к тому, что эксперимент становится одним из основных средств получения информации о проектируемом или исследуемом процессе или явлении.

Приоритет статистических методов в познании окружающего мира (на данном этапе развития) отразился и в изменении парадигмы. Парадигма — общепризнанная непротиворечивая система взглядов, гипотез, теорий существующая на определенном этапе развития науки и используемая для получения научно обоснованных выводов и рекомендаций. Другими словами парадигма — это некоторая модель, взятая в одной из наиболее

[4] Возьмем, например, закон Ома: хорошо известно, что он получен на основании обобщения сотен опытов. Мы знаем, что он не выполняется при сверхнизких (разных для различных проводников) температурах. Он выполняется лишь приблизительно. И чем точнее и тщательнее мы будем проводить измерения, тем больше будет видно, что результат каждого измерения — случайная величина. Один из великих физиков нашего времени говорил, что если бы физики прошлого имели бы такие точные инструменты, как мы, то они не открыли бы ни одного физического закона.

развитых на данный момент наук и считается наиболее подходящей для описания явлений и процессов во всех других науках. Долгие столетия такие модели заимствовались из физики. Это была механическая парадигма, затем гидравлическая и электрическая. Последние десятилетия эта роль перешла к кибернетике. В качестве парадигмы в общем случае выступает стохастический автомат или же более простая статистическая модель [25].

Без применения статистических методов невозможно решить основную задачу, которая возникает перед каждым экспериментатором: как провести и обработать результаты эксперимента, чтобы при заданных ограничениях на затраты получить максимум достоверной информации.

Таким образом, в современных условиях статистические методы стали одним из главных инструментов не только в научном познании, но и в технологических и конструкторских разработках. Анализ распространенности применения статистических методов подтверждает вышесказанное[5].

1.1. Особенности современного этапа развития и использования экспериментально-статистических методов

Существуют следующие виды лжи: ложь, наглая ложь, предвыборные обещания и статистика.

Дизраэли

Гораздо труднее увидеть проблему, чем найти ее решение.

Джон Д. Бернал

Несмотря на широкое распространение и успех применения статистических методов, мы не должны скрывать целый ряд проблем,

[5] Один из американских журналов опубликовал итоги исследования, проведенного в 500 крупнейших фирмах США. Вот они: оказывается, 98,4% фирм эпизодически или постоянно применяют методы математико-статистического анализа, машинную иммитацию — 87,1%, сетевое планирование — 74,2%, линейное программирование — тоже 74,2%, методы теории очередей — 59,7%, нелинейного программирования — 46.8%, динамического программирования — 29,7%; наименьшее применение получили пока методы теории игр — 20,6%. Иными словами, по многим направлениям — свыше трех четвертей фирм применяют эти методы!.. 87,1 % фирм полагает, что затраты на их применение окупаются полученными доходами. (Взято из: Лопатников Л.И. Популярный экономико-математический словарь.— 3-е изд., дополненное.— М.: Знание, 1990.— 256 с.) В Японии за один лишь 1987 год зарегистрировано 1 млн применений только метода Тагути (не говоря уже обо всем остальном), что дало общий экономический эффект 7 млрд долларов США.

возникших как раз в процессе их экспансии. Это как проблемы применения методов, так и проблемы их внутреннего развития. Корень всех возникших затруднений состоит в том, что корректное применение любого статистического метода или критерия связано с выполнением некоторого набора допущений и предпосылок. Если они не выполняются, то правильных выводов из применения метода сделать, строго говоря, нельзя. Допущения и предпосылки вводятся статистиками для того, чтобы можно было теоретически обосновать получаемое решение.

Для регрессионного анализа обычно выдвигаются следующие предпосылки и допущения.

1. Математическое ожидание случайной ошибки в каждом опыте равно нулю $E(\varepsilon_n) = 0$.
2. Случайные ошибки в любых двух опытах независимы между собой $cov(\varepsilon_u, \varepsilon_l)=0$ при $u \neq l$ и $u, l = 1, 2, \ldots n$.
3. Дисперсия случайной ошибки не изменяется $\sigma^2(\varepsilon_u) = \sigma^2 = const$, $u=1, 2, \ldots N$.
4. Случайная ошибка распределена по нормальному закону.
5. Матрица, по которой выполняется расчет коэффициентов регрессии не случайная.
6. Ранг матрицы, по которой выполняется расчет коэффициентов регрессии, больше или равен числу коэффициентов $rank(F) \geq k$.
7. На коэффициенты регрессии не накладывается никаких ограничений.

Только при выполнении перечисленных условий оценки коэффициентов регрессии будут эффективными (имеют наименьшие дисперсии среди всех несмещенных оценок), состоятельными (при увеличении числа опытов оценка по вероятности сходится к истинному значению) и несмещенными (математическое ожидание коэффициентов равно истинному их значению).

Мы обращаем внимание на наиболее важные вопросы при использовании регрессионного анализа, без учета которых возможны наиболее тяжелые последствия. Таким вопросом является выбор матрицы независимых переменных. Если столбцы этой матрицы имеют сильную корреляционную связь, то это является нарушением, которое может привести к следующим последствиям:

— невозможность корректного применения t и F статистических критериев;

— невозможность разделить влияние различных эффектов (каждый коэффициент регрессии содержит смешение оценок взаимозависимых регрессоров);

— при сильной закоррелированности эффектов матрица $(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}$ становится плохо обусловленной, а вследствие этого оценки коэффициентов регрессии неустойчивыми.

Таким образом, может сложится такая ситуация, при которой

полученная модель и все ее статистические оценки не имеют никакого смысла.

Как правило, выполнение проверок и корректировок требует привлечения дополнительных ресурсов на эксперимент или проведения специальных расчетов которые соизмеримы (а часто и превосходят) с затратами на собственно регрессионный анализ. Специального программного обеспечения для выполнения этих действий не существует. Кроме того, при некоторых нарушениях, несмотря на затраченные усилия, качество полученной модели будет неудовлетворительным с точки зрения инженера.

В частности, дисперсии оценок коэффициентов регрессии становятся очень большими, а сами коэффициенты статистически незначимыми; построить доверительный интервал для каждой из них невозможно; статистические оценки становятся закоррелированными и не отражают истинного соотношения между регрессорами; оценки становятся неустойчивыми по отношению к малым изменениям исходных данных выборки.

Но, кроме этих, достаточно неприятных последствий высокой закоррелированности существуют и другие, не менее неприятная сторона.

Оказывается, что при неправильном выборе частной функции регрессии оценки коэффициентов регрессии будут несмещенными, состоятельными и эффективными только при отсутствии закоррелированности[6] эффектов.

Кроме того, при разработке и применении метода случайного баланса было замечено, что закорелированность матрицы условий эксперимента сильно влияет на возможность правильного выделения значимых регрессоров[7]**.** К сожалению, в реальных условиях производства зачастую невозможно проверить выполнение всех предпосылок и допущений. А задачу-то решать надо! В связи с этим в настоящее время происходит переход от математической статистики к робастной, непараметрической статистике и анализу данных. В непараметрической статистике среди предпосылок нет фиксации определенного закона распределения, а в робастной допускаются умеренные отклонения от него. При этом средством

---

[6] Демиденко Е.З. Линейная и нелинейная регрессии.— М.: Финансы и статистика, 1981.— 302 с.

[7] Барский В.Д., Забенко Л.А., Аксенина А.А., Беднов В.М. К вопросу о построении матрицы планирования отсеивающего эксперимента. // Заводская лаборатория.— 1971.— № 7 — Т. 37.— С. 721-825.; Лапина З.С., Слободчикова Р.И. Исследование границ применимости алгоритма случайного баланса. // Заводская лаборатория. — 1971.— Т. 37.— № 7.— С. 818-821; Слободчикова Р.И., Фрейдлина В.Л., Лапина З.С., Налимов В.В. Повышение эффективности метода случайного баланса путем применения ветвящейся стратегии и электронно-вычислительных машин // Заводская лаборатория.— 1966.— № 1.— Т. 32.— С. 53-58.

обоснования методов служит не аналитическая теория, а вычислительный эксперимент. Робастная статистика разрабатывает такие методы, которые дают удовлетворительные результаты при нарушении предпосылок и допущений. Анализ данных ориентирован на решение определенных классов прикладных задач. При этом процедуры обработки скорее носят логико-эвристический характер, чем статистический. Теоретически обосновать такие методы в большинстве случаев трудно или невозможно. Поэтому для проверки приянтых рекомендаций производится вычислительный эксперимент. При этом создаются в большом количестве искусственные данные, которые имитируют различные условия экспериментов и на этапе обработки данных проверяется эффективность предложенного метода. В зарубежных научных изданиях правильно проведенный вычислительный эксперимент считается достаточным обоснованием разработанного метода.

С экспансией методов статистики и невозможностью проверки допущений и предпосылок связана еще одна серьезная проблема в применении статистических методов. Эти методы применяются для решения все более сложных прикладных задач. При этом в состав пользователей вовлекается большое число людей, не являющихся специалистами в области статистики. В результате этого возрастает число неудач в применении экспериментально-статистических методов, а вследствие этого разочарование в их эффективности и отказ от использования.

Эти проблемы связаны с двумя принципиальными моментами: во-первых, применение любого статистического метода требует выполнения ряда предпосылок, которые в реальных условиях во многих случаях невозможно проверить; во-вторых, проблема формализации, в результате которой зачастую происходит не решение прикладной задачи, а применение статистических методов. Это приводит к получению многовариантных и противоречивых результатов, или к результатам, степень достоверности которых неизвестна.

В разных странах к решению этих проблем подходят по разному. В США и европейских странах упор делается на подготовку относительно большого числа высококвалифицированных специалистов. В Японии разрабатывают относительно простые приемы (и даже градации приемов для специалистов разной квалификации) для массового применения. Там выделено несколько классов прикладных задач, которые решают большую часть (по количеству) потребностей и для каждого класса разработана достаточно простая и эффективная технология решения. Оба эти пути снимают напряженность, но не решают проблему в принципе, поскольку высококвалифицированный специалист смотрит на проблему со стороны статистики, а японский подход слишком узкоспециализирован.

В бывшем СССР положение было весьма своеобразным. Специалистов

было относительно мало, подавляющее большинство работало в теоретической области и к практическому применению статистических методов никакого отношения не имели. Пик распространения экспериментально-статистических методов имел место в семидесятые годы, а затем ввиду уже описанных причин и отсутствия нормальной функционирующей экономики сошел на нет. Тем не менее, проявилась тенденция к созданию информационных технологий и квазиинтеллектуального программного обеспечения, реализующего определенные классы прикладных задач[8].

Рассмотрим один из широко применяемых методов — регрессионный анализ. Одна из нерешенных проблем статистики — это определение частной структуры уравнения регрессии или как стыдливо говорят статистики — определение наилучшего подмножества регрессоров. Проблема здесь принципиальная — те, кто пользуются статистикой считают, что уравнение регрессии должно отражать «истинную» структуру связей в изучаемом процессе или системе и быть пригодным для качественного анализа происходящих явлений. Статистики (подавляющее большинство) говорят лишь о некотором наборе статистических свойств и слышать не хотят о какой-то там истинной структуре.

Для иллюстрации абсурда этой ситуации возьмем классический пример, известный практически всем, кто пользовался или изучал регрессионный анализ — регрессию Лонгли.

Это регрессия от шести факторов: y — общее число занятых в экономике США (тыс. чел.); X_1 — дефлятор (индекс) цен (%); X_2 — валовой национальный продукт (млрд дол.); X_3 — общее число безработных (тыс. чел.); X_4 — число военнослужащих (тыс. чел.); X_5 — неработающее население от 14 лет (тыс. чел.); X_6 — год. Данные были взяты за период с 1947 по 1962 год. Оказалось, что матрица данных плохо обусловлена (число обусловленности *cond* = $4{,}8 \times 10^{59}$) и при расчете по разным программам и на разных ЭВМ оценки коэффициентов регрессии получались не только с разными значениями, но и с разными знаками. С помощью специального счетного устройства были выполнены расчеты с 40 значащими цифрами и получена следующая модель.

$$y = 15{,}0619X_1 - 0{,}0358X_2 - 2{,}0202X_3 - 1{,}0332X_4 - 0{,}0511X_5 + 1829{,}15X_6 - 3482258{,}635$$

Эта регрессия до настоящего времени применяется для проверки точности работы программ по регрессионному анализу. И это несмотря на то, что некоторые теоретики показали её бессмысленность (A.E. Beaton, D.B. Rubin, J.L. Barone в 1976 и Е.З. Демиденко в 1981 г.), не предлагая, правда, другого решения.

[8] Это система СИТО (В.В. Александров, А.И.Алексеев, Н.Д. Горский — г. Санкт-Петербург), ПС «Чегет» (М.В. Еханин, Л.П. Рузинов, Р.И. Слободчикова,— г. Москва), ПС «ПРИАМ» (С.Н. Лапач, С.Г. Радченко, П.Н. Бабич — г. Киев).

Это задача решалась с помощью ПС ПРИАМ в автоматическом режиме (т.е. структура уравнения регрессии формировалась программой). При этом получена модель

$y = 65317{,}00+5810{,}429x_2$,

где $x_2 = 0{,}00598102 *(X_2 – 387{,}698)$;

$R = 0{,}983552$, $\gamma = 5$ для критерия Бокса-Веца, *cond* = 1.

По этой модели погрешность аппроксимации находится в пределах 0,4…1,2 %, а по модели Лонгли — 13,5 … 28,5 %.

Этот пример ясно показывает, что нужно решать задачу, а не применять методы, поскольку мы получаем ответ на вопрос, который не задавали, а наш вопрос остается без ответа.

Разработанная технология ориентирована на массовое решение задач, связанных с построением математических моделей по результатам эксперимента и их использование. Технология полностью обеспечивает всю схему решения от формализации до применения и окончательных выводов. Технология вместе с ПС ПРИАМ (“Планирование, Регрессия, и Анализ моделей”) обеспечивает пользователю, не имеющему глубоких знаний в области статистики, быстрое получение высокоэффективного результата. При этом он получает максимум достоверной и надежной информации при заданных ограничениях на ресурсы и гарантирован от неправильного применения методов и связанных с этим ошибок в выводах.

Данная работа ориентирована на специалистов в конкретной области знаний, не имеющих подготовки в области статистики, но которые вынуждены применять экспериментально-статистические методы в своей работе. Это могут быть научные работники, инженеры, аспиранты, студенты. В нем излагается схема решения прикладной задачи по принципу: что должен делать пользователь, чтобы получить хороший результат. В наиболее важных местах приводится объяснение принимаемых решений. Именно объяснение, а не доказательство, т.к. это руководство к работе, а не теоретическая монография. По этой же причине не приводится описание алгоритмов.

Эффективность технологии подтверждена успешным применением с 1980 года для решения задач на нескольких десятках предприятий и организаций Украины и России. Программное средство ПРИАМ, поддерживающее данную технологию продается с 1982 года для ЕС ЭВМ, с 1991 года для ПЭВМ.

1.2. Общая схема решения задачи

Незаурядный моряк использует свою незаурядную рассудительность, чтобы избежать ситуаций, требующих его незаурядного мастерства.

Ричард А. Кейхилл

«Столкновения судов и их причины»

В общем случае решение прикладной задачи проходит следующие этапы (в упрощенном виде):

1. Формализация задачи.
2. Конструирование плана эксперимента.
3. Проведение экспериментов.
4. Предварительный статистический анализ результатов эксперимента.
5. Построение математических моделей по результатам эксперимента.
6. Анализ качества полученных моделей.
7. Проведение вычислительного эксперимента с использованием построенных моделей.
8. Формирование выводов, рекомендаций и отчета о проведенном исследовании.

Часть этапов может отсутствовать, возможны возвращения на предыдущие этапы для уточнения постановки задачи, проведения дополнительных экспериментов.

Фактически указанная цепочка является технологическим процессом получения и обработки информации при решении задач моделирования и оптимизации. От организации этого процесса и применяемых средств в значительной степени зависит конечный результат: качество (достоверность) информации, количество (т.е. объем и полнота) и себестоимость (т.е. размер затрат). Целью организации такого процесса является получение максимального количества достоверной информации при ограничении сверху на используемые для этого ресурсы.

Особенностью данного технологического процесса получения информации является то, что при решении сложных задач неэффективное выполнение ранних этапов невозможно исправить даже за счет применения самых совершенных и мощных средств на последующих этапах. Поэтому на каждом этапе необходимо применение наиболее эффективных, с точки зрения получения конечного результата, методов и средств их реализации.

Следует четко уяснить себе, что это технологический процесс получения информации. Этот процесс многостадийный, наукоемкий и сложный. Нарушения в выполнении рекомендаций получения информации неизбежно приведут к потере качества или даже к полной неудаче.

Рассмотрим кратко содержание этапов.

Формализация задачи по сути означает формулировку прикладной задачи в виде, пригодном для решения экспериментально-статистическими методами. Это наиболее ответственный этап. От правильности его выполнения напрямую зависит конечный результат: на неправильный вопрос вы никогда не получите ожидаемого ответа. В результате формализации мы формулируем цель, определяем требуемые ресурсы и условия проведения исследования.

После проведения формализации выполняется конструирование плана

эксперимента, представляющего собой описание экспериментов, которые нужно выполнить для решения задачи. От плана эксперимента зависит объем информации, которую вы получите, ее качество и надежность, а также усилия, которые необходимо затратить на ее получение. Неправильный выбор плана может привести к тому, что только профессионалу, и то с большим трудом да и не всегда, удастся получить информацию из ваших данных.

При проведении экспериментов необходимо следить за тщательным их выполнением не только с точки зрения прикладной области, но и за выполнением требований статистики (разбиение плана эксперимента на блоки, рандомизация опытов и т.д.), т.к. невыполнение этих требований может привести к внесению искажений, от которых трудно избавиться.

После проведения экспериментов проводят предварительный статистический анализ, который позволяет оценить уровень «шума» и отбросить грубые ошибки — «выбросы».

На следующем этапе определяется структура уравнения регрессии и рассчитываются оценки коэффициентов регрессии и их статистические характеристики.

После этого необходимо провести анализ статистических и потребительских свойств модели с целью определения пригодности ее для использования. Обычно анализируются такие свойства, как информативность, адекватность, устойчивость, а также описывающие и предсказующие свойства модели. После анализа принимается решение о ее пригодности или о необходимости доработки.

Этап вычислительного эксперимента представляет собой уже непосредственное использование модели для изучения объекта или процесса. При этом исследуемый объект заменяется построенной моделью. Обычно на этом этапе выполняются графические исследования — семейства частных уравнений и линии равного уровня, поиск оптимальных условий, предсказание значений.

На последнем этапе формулируются выводы, рекомендации и заключения.

В данной работе, в тех случаях, когда говорится об использовании для решения определенных программных средств, упоминаются мало известные в широких кругах ПРИАМ, DESFACT, OptimeChoice (распространены не более чем в нескольких десятках экземпляров). Это связано с тем, что по ряду параметров они являются уникальными и не имеют близких по возможностям аналогов среди существующих пакетов: ПРИАМ по построению регрессионных моделей, DESFACT по конструированию многофакторных регулярных планов, OptimeChoice — многокритериальной оптимизации. Это связано со следующими причинами:

- «западные» программные средства создаются не специалистами по статистическим методам (исключение BMDP). Крупные

специалисты по статистике не привлекаются к их созданию, т.к. это повысило бы стоимость ПС до неприемлемого уровня. В связи с этим ПС обычно реализуют стандартные, классические процедуры обработки данных не ориентируясь на прикладную сторону решаемых задач;

- использование пакетов предполагает достаточно высокий уровень знакомства с используемыми методами, способность самим выбрать последовательность действий, определить корректность применяемых методов к конкретным данным и оценить правильность полученных результатов.

Программные средства, на которые мы ссылаемся реализуют оригинальные передовые технологии и не требуют от пользователя глубоких знаний в области используемых методов, поскольку сами организовывают процесс решения, обеспечивая надежность и корректность полученных результатов. Для специалистов остается возможность самому организовывать вычислительный процесс и его параметры по своему усмотрению.

Особенностью описываемой технологии является то, что в ней удалось избавиться в значительной степени от двух основных недостатков классической теории планирования экспериментов и регрессионного анализа, которые являются тормозом широкого применения указанных средств — планирования и обработки модели заранее заданной структуры и обязательного использования стандартной области планирования.

Как известно, классическая процедура построения плана предполагает заданную структуру уравнения регрессии, которая практически никогда не бывает известной. Кроме того, каждый из таких планов рассчитан на определенную сложность модели. В случае, если возможности плана эксперимента оказываются недостаточными, необходим новый план, а ресурсы на проведение экспериментов становятся затраченными впустую. Хотя в ряде литературных источников указывается на недостатки существующих планов, но альтернативы им не предложено.

В описании технологии обработки данных предлагается использование квазиробастных планов: имеется в виду робастность по Хьюберу, т.е. план оптимален независимо от структуры модели.

Как уже упоминалось, технология обработки данных с использованием ПС ПРИАМ позволяет получить регрессионные модели при нестандартной области планирования. Под нестандартной областью планирования понимается область, не представляющая собой гиперпараллелепипед, сферу, симплекс, что связано с мультиколлинеарностью факторов. Причина мультиколлинеарности факторов обычно связана с физической или статистической зависимостью их друг относительно друга.

Решение регрессионных задач в условиях мультиколлинеарности относится к классу некорректно поставленных задач, решение которых

неустойчиво к малым изменениям входных переменных.

Применение технологии обработки данных с использованием ПС ПРИАМ позволяет обеспечить при комплексном подходе высокую надежность конечного результата и экономию ресурсов, необходимых для исследования. Его использование означает по сути переход к индустриальным методам решения задач — решения задач «на потоке», обеспечивая высокую производительность и качество результатов.

1.3. Случайные величины

Статистика — совокупность методов, которые дают нам возможность принимать решение в условиях неопределенности.

Абрахам Вальд

1.3.1. Шкалы измерений

Измеряй все доступное измерению и делай недоступное измерению доступным.

Галилео Галилей

Обработать статистическими методами возможно лишь то, что можно измерить. В связи с этим необходимо рассмотреть шкалы измерений. Существуют следующие шкалы измерений:

- ❑ шкала классификации (наименований);
- ❑ шкала порядка;
- ❑ шкала интервалов;
- ❑ шкала отношений.

Рассмотрим особенности этих шкал.

Шкала классификации (наименования, номинальная). Никакие операции по сравнению невозможны, кроме «равны» и «не равны». Нумерация или поименование служит лишь для идентификации объекта — номер дома, номер на майке спортсмена и т.п.

Шкала порядка. Возможно сравнение объектов по величине (больше или меньше). Другие операции невозможны. Примером может служить шкала твердости минералов, в которой имеется ряд эталонных минералов, выстроенных в ряд: каждый последующий минерал в котором тверже предыдущего. Возможны только операции сравнения по типу больше, меньше, равно.

Шкала интервалов. В этой шкале возможно не только сравнение по величине, но и определение насколько больше (т.е. возможны операции сложения и вычитания). Примером могут служить шкалы измерения температуры (Цельсия, Кельвина, Фаренгейта, Реомюра).

Шкала отношений. В этой шкале возможны все операции (сравнения, сложения и вычитания, умножения и деления), т.е. возможно выяснение вопроса: во сколько раз. Пример — вес, длина и пр. В этих случаях существует естественная точка отсчета.

Следует сказать, что в процессе развития соответствующих наук и средств измерения возможен переход от одной шкалы измерений к другой, более совершенной. Так, например, первые термометры измеряли температуру в шкале порядка (умерено, тепло, горячо и т.п.).

Иногда говорят также о дискретных и непрерывных шкалах измерений. В общем случае к дискретным относятся шкала классификации и шкала порядка.

1.3.2. Закономерность и случайность

Все, что тебе представляется случайным стечением обстоятельств, вовсе таковым не является.

Станислав Лем. «Собысча»

При изучении предметов в школе, а зачастую и в высших учебных заведениях, неявно предполагается детерминированность, т.е.: каждое событие является следствием другого; физические законы представляют собой строгие закономерности зависимости одних величин от других. Вместе с тем повседневная жизнь и работа каждого человека постоянно опровергают это положение. Так при проверке любых физических законов даже на уровне лабораторных занятий в школе или вузе, обнаруживается, что каждое новое экспериментальное определение величин дает немного отличающиеся