Я. И. ХУРГИН

6 Ф 6. 5
Х 98
УДК 62-50

Да, нет или может быть. Хургин Я. И. Главная редакция физико-математической литературы издательства «Наука», М., 1977, 208 стр.

В книге рассказывается о статистической теории управления и эксперимента — научном направлении, находящемся на стыке нескольких наук: теории управления, математической статистики и теории эксперимента. На примерах из разных областей науки и техники в доступной форме разъясняются принципиальные положения статистической теории управления и эксперимента, *освещаются современные принципы* построения математических моделей и обработки результатов наблюдений, статистической проверки гипотез и планирования экспериментов. Книга адресована широкому кругу читателей, интересующихся проблемами прикладной математики, в том числе специалистам, так или иначе связанным с экспериментом, но не имеющим достаточной математической подготовки, особенно в области теории вероятностей и ее применений (инженеры, геологи, химики, биологи, медики, экономисты).

Автор книги, будучи специалистом в области прикладной математики, кибернетики, применения математических методов в различных областях науки и техники, широко известен также как популяризатор науки. Его книга «Ну и что?», дважды издававшаяся в серии «Эврика» (издательство «Молодая Гвардия»), хорошо знакома любителям математики.

Табл. 10, илл. 67.

Художник К. Р. БОРИСОВ

Х $\frac{30501-144}{053(02)-77}$ 167-77

Зачем и кому написана эта книжка

Как-то в «Литературной газете» была полемика между учеными о том, нужна или нет, полезна или вредна популярная литература о науке. Сама постановка вопроса меня крайне удивила. Почему-то никому не приходит в голову обсуждать полезность информации о всех странах и народах мира, и телепередача «Клуб кинопутешествий», которую интересно ведет Юрий Сенкевич, — это любимая передача большинства моих знакомых. Но ведь и жизнь огромного мира науки — это все та же жизнь окружающего нас мира, и чем больше человечество о ней будет знать, тем лучше. Конечно, рассказ о науке, особенно теоретической, внешне менее выгрышен, чем о племени бушменов или о дворцах Франции, но высшее достижение человечества — игра идей не менее прекрасна. Познать жизнь любого народа мешает прежде всего языковый барьер, и поэтому весь мир изучает иностранные языки. «Чужая» наука непонятна тоже из-за языкового барьера, но когда пробьешься через его густой кустарник, то выйдешь к чистому роднику идей, понятных каждому.

Один из моих школьных учителей говорил: «Нет плохих слов — есть плохие уста». Перефразируя, скажу: «Нет непонятных идей в науке — есть нежелание сделать их доступными». Конечно, рассказать неспециалисту о достижениях молекулярной биологии или астрофизики нелегко. Но теория происхождения видов Дарвина тоже

многим казалась сначала непонятной, а теперь ее знает каждый школьник.

Мне — математику в жизни повезло: пришлось знакомиться с радиотехникой и физиологией, кибернетикой и психиатрией, теорией передачи информации и технологией переработки нефти, теорией управления и геофизикой. Каждая из этих различных областей науки имеет, как народность, свой язык, и далеко не сразу становится ясным, что радиоимпульс и спайк нейрона, сейсмограмма и отклик системы управления — все это почти одно и то же. Разобравшись в основах разных наук в конце концов понимаешь, как в них много общего, значительно больше, чем кажется на первый взгляд.

Основные успехи сейчас достигаются на стыках наук, и поэтому борьба за воспитание научных кадров высшей квалификации — кандидатов и докторов наук — как узких специалистов мне представляется не отвечающей духу времени. Наоборот, разные, представляющиеся совсем далекими науки могут очень активно обогащать друг друга. Именно поэтому я и написал книжку, поставив себе целью рассказать популярно, понятно о бурно развивающемся в последние два десятилетия научном направлении, которое кратко можно назвать статистической теорией управления и эксперимента.

Когда берешься за написание популярной книжки о каком-либо разделе науки, то возникает вопрос об отборе материала. Поэтому прежде всего я снял с полок довольно много толстых и тонких книг с похожими, а то и совпадающими названиями и обнаружил совершенно разное содержание — уж очень обширна область науки, о которой идет речь. Конечно, безнадежно было браться за рассказ обо всех интересных идеях или методах. Поэтому пришлось в отборе материала, да и в изложении в основном отказаться от ссылки на коллег и полагаться на свой вкус. Я старался донести до вас, читатель, идеи, освободив по возможности изложение от тумана терминологии и груза деталей. Не всюду удалось избавиться от терминологии в равной мере, но, быть может, это и не беда: если после прочтения книжки вы заинтересуетесь обсуждаемыми проблемами и возьметесь изучать их всерьез по учебникам и монографиям, то такое предварительное знакомство с идеями и терминологией облегчит вам жизнь.

Почему же так трудно было отобрать материал? Вы, конечно, представляете себе, что такое управление, иначе

едва ли взяли бы в руки эту книжку. Огромная литература посвящена проблемам управления без всяких предположений о наличии случайных воздействий на управляемый объект, будь то самолет или прокатный стан, завод или государство, живой организм или просто некоторый абстрактный объект. Но возможно ли описать систему управления самолетом без учета нерегулярностей плотности атмосферы, постоянно меняющейся силы ветра, малых механических неоднородностей в конструкции, несимметрии распределения его груза, без учета движения людей внутри салона во время полета и многого другого? Можно ли описать управление жизнедеятельностью инфузории или слона без учета воздействия внешней среды, изменяющейся все время, но отнюдь не регулярно; все те случайные удачи и неприятности, на которые и слон, и инфузория должны как-то реагировать ежедневно, ежечасно, ежесекундно? Сомнительна подобная возможность, не так ли?

Но не следует ли из наших рассуждений невозможность вообще описать какую угодно систему управления без учета случайности? Нет, не следует, и вот почему.

Вы привыкли к изменениям продолжительности дня и ночи: от 22 июня до 22 декабря продолжительность дня уменьшается, а затем увеличивается. Все это совершенно регулярно, точно предсказуемо на годы и десятилетия вперед. Но представьте себе нашу жизнь в ситуации, когда продолжительность суток и смена дня и ночи были бы случайны, то есть если бы скорость вращения Земли вокруг своей оси не была постоянна, а изменялась по воле случая, как настроение четырнадцатилетней девушки: то она уныло плетется, то резво бежит, и эти перемены совершенно не регулярны. Вдруг быстро взошло Солнце, и надо спешно собираться на работу, но день тянется и тянется, — скорость вращения Земли внезапно упала. Затем скорость неожиданно возросла, и вы не успеваете на свидание, назначенное перед заходом Солнца. Вечер пролетел быстро, но выспаться невозможно — ночь кончилась за три часа. Потом короткий день, и вы не успеваете даже пообедать, как наступает ночь, на этот раз длинная... Все рефлексы сна и бодрствования нарушены. Да, при наших привычках такой жизни не позавидуешь.

Земля вращается вокруг оси и на самом деле не совсем регулярно, ее движение подвержено случайным воздействиям — то метеорит ударит, то пролетит вблизи комета,

но эти воздействия маленькие, и на продолжительности суток, смене дня и ночи они отражаются весьма незначительно, и мы при решении повседневных задач можем спокойно ими пренебречь и рассматривать движение Земли вокруг оси как совершенно регулярное.

Температура воздуха подвержена значительно более ощутимым случайным изменениям. Но мы к ним приспособились, непрерывно управляя поведением своим, своих детей или подчиненных: то надеваем теплое пальто, то берем зонтик, открываем или закрываем окна, топим печи или чиним крышу. А как легко и вольготно нам жилось бы, если бы температура воздуха на Земле изменялась, скажем, уменьшаясь от $+25^\circ$ С в день летнего солнцестояния до -25° С в день зимнего солнцестояния, а затем бы обратно возрастала по определенному регулярному закону, например по синусоиде. Никаких забот о неожиданном похолодании, заранее известно, когда какой наряд надевать, к какому числу надо уже сшить шубу или приобрести новый купальник, когда закрывать катки и открывать футбольный сезон, не нужно летом топить дачу или хлюпать по снежной каше в январскую оттепель. Таким образом, вопрос о том, следует ли учитывать случайные воздействия при решении любой задачи, в том числе задачи управления, зависит от условий, и иногда можно пренебречь случайными воздействиями, а в других ситуациях нельзя. Однако ситуаций последнего сорта видимо-невидимо, и именно о них в книге будет речь.

Случайные воздействия, возмущения, силы естественно описывать, опираясь на разделы науки, содержание которых — изучение событий, величин, процессов случайных, то есть с позиций теории вероятностей и математической статистики.

Хотя я и анонсировал доступность изложения, все же для сознательного чтения книги полезно иметь элементарные сведения по теории вероятностей. Сейчас начала теории вероятностей изучают не только в вузовском курсе высшей математики большинства специальностей технического, естественного и экономического профиля, но они включены в новую программу средней школы. Поэтому я и адресую книгу широкому кругу читателей.

Все же следует учитывать реальное положение дел: когда сдают экзамен или зачет, не имея в виду в ближайшем будущем активно использовать изучаемый материал, то его доносят до экзаменатора на цыпочках, боясь расплескать, но, сдав, после вздоха облегчения почти сразу

забывают. Поэтому кое-какие основные идеи и понятия теории вероятностей будут вкратце повторены по ходу дела.

Честно говоря, для написания этой книжки был еще один стимул. Десять лет назад была издана моя популярная книжка «Ну и что?» (Издательство «Молодая Гвардия», серия «Эврика», 1967 г.). Эта книга очерков о математике в широком смысле, о ее методах и идеях и, главное, о связи математики с другими науками, о том, как математик может сотрудничать, совместно и плодотворно работать с научными сотрудниками других специальностей. Судя по откликам читателей, она принесла пользу специалистам — биологам, инженерам, экономистам, технологам, химикам!

Предлагаемая вам сейчас книга носит несколько иной характер. Она не сборник отдельных новелл, а посвящена хотя и широкому, но определенному научному направлению. Однако и управление, и эксперимент занимают весьма большое место в жизни, и мне хотелось бы, чтобы многим специалистам — нематематикам принесла пользу и эта книга.

Друзья стимулировали меня к работе над этой книжкой. Критические замечания и советы Л. Н. Большева, В. М. Глоговского, Я. А. Когана способствовали улучшению содержания и изложения, и даже моя дочь Ирина отредактировала часть текста, где ни о какой математике нет речи. А. А. Молявко практически безошибочно и с неизменным доброжелательством неоднократно перепечатывала разные варианты. Всем им я приношу свою признательность.

Москва, июнь 1977 г. *Я. Хургин*

Неопределенность и случайность

До сих пор в книгах, адресованных читателю — неспециалисту и использующих понятия теории вероятностей и математической статистики, принято приводить в начале или в приложении раздел, посвященный элементам теории вероятностей, с определениями таких понятий, как вероятность, условная вероятность, распределение вероятностей, случайная величина, математическое ожидание, дисперсия и т. д. Мне не хочется следовать этой традиции, ибо изложение вероятностных основ на нескольких страницах создает у читателя иллюзию, будто он уже освоил эти понятия, в то время как именно понимание исходных посылок и основных положений теории вероятностей и математической статистики, вопросов применимости теории, ее особенностей и возникающих парадоксов наталкивается на значительные психологические трудности особенно у лиц, вышедших из студенческого возраста. В то же время формальная сторона — математический аппарат теории вероятностей и математической статистики — ничем, по существу, не отличается от математического анализа и линейной алгебры и дополнительных трудностей не вызывает. В кратких учебных пособиях обычно мало внимания уделяется тем исходным посылкам, на которых строится теория вероятностей, и о них я кое-что расскажу.

Вы выходите из дому и первой встречаете блондинку. Нет, не ту блондинку, которая на вас произвела сильное

впечатление позавчера, а вообще какую-нибудь блондинку, то есть не рыжую, брюнетку или шатенку. С позиций изучаемой области науки появление блондинки — событие, оно может произойти или не произойти, и ни о чем другом пока речь не идет. В то же время с обычных житейских позиций встреча с блондинкой может быть событием значительным, неприятным или не произвести никакого впечатления, но о важности событий пойдет речь в разделе, посвященном теории риска.

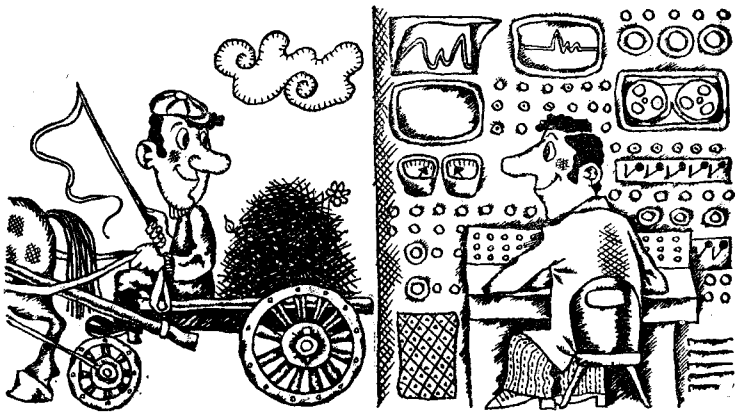
Если всякий раз, когда вы выходите из дому, регистрировать, встречаете ли вы первой блондинку или нет, то можно подсчитать частоту наступления события *первое встреченное лицо — блондинка* (частота — это отношение числа раз, когда вы встретили блондинку к общему числу наблюдений). Прежде всего следует обратить внимание на реальную возможность проводить наблюдения многократно в одинаковых условиях. Будет то лето или зима, солнечный день или дождливый вечер, шансы встретить первой блондинку при вашем выходе из дома в общем-то одни и те же.

Мы здесь предполагаем возможность производить наблюдения в одинаковых условиях не только многократно, но неограниченно, сколько угодно раз. Весьма правдоподобно, что с увеличением числа наблюдений частота наступления изучаемого события будет мало и несистематически меняться, и если производить одинаковое количество наблюдений в разные месяцы, то частота наступления изучаемого события также будет мало колебаться. Пусть число наблюдений велико, и вы выбираете заранее какое-либо подмножество наблюдений, например каждое третье или первые сто пятнадцать в каждой тысяче, или как вам еще придет в голову, лишь бы число наблюдений в подмножестве было достаточно велико и неограниченно возрастало вместе с общим количеством наблюдений. Тогда нужно, чтобы частоты наступлений изучаемого события, вычисленные по всем наблюдениям и по отобранному подмножеству, были близки. Если изучаемое событие обладает такими свойствами, то оно называется с л у ч а й н ы м. Именно в этих условиях вводится понятие вероятности наступления события, как обобщение частоты его наступления, причем вводится аксиоматически.

Исход олимпийских игр, скажем, по спортивной гимнастике, завоевавшей интерес широкого зрителя, заранее предсказать нельзя, и, конечно, многое зависит от случая.

Как мы недавно наблюдали, могут развалиться брусья во время выступления претендентки на первое место. Могут быть неожиданные травмы, болезни, кто-то находится не в лучшей форме... И все же с позиций теории вероятностей исход олимпийских игр не есть случайное событие: нельзя повторять олимпийские игры многократно в одних и тех же условиях. На следующих играх будут другие участники, другое место проведения игр и т. д. Такие события, в отличие от случайных, называют неопределенными. С математических позиций они изучаются в теории игр. Но в этой книжке будет рассказано главным образом о случайных событиях, и поэтому основным аппаратом служит теория вероятностей и математическая статистика.

Итак, я хочу подчеркнуть еще раз: не всякое событие, исход которого неизвестен и непредсказуем однозначно, называется случайным в используемом нами языке. Событие становится случайным при наличии определенных условий, ограничений, о которых только что шла речь, — я их буду называть статистической устойчивостью. И обсуждаемая далее теория относится не к чему угодно, лишь бы оно было заранее неизвестно, а к событиям случайным, статистически устойчивым.



Управление

Вы, конечно, читали знаменитый роман Жюля Верна «Дети капитана Гранта». Но мне хочется напомнить его основные вехи.

Летом 1864 года вблизи острова Арран мчалась на всех парах великолепная яхта «Дункан», завершающая пробное плавание.

Хозяин яхты лорд Эдуард Гленарван, один из шестнадцати шотландских пэров, заседающих в палате лордов, совершал это путешествие вместе со своей молодой и очаровательной женой Элен, двоюродным братом майором Мак-Наббсом и капитаном Джоном Манглсом.

Вахтенный матрос заметил за кормой огромную рыбу-молот. Яхту остановили, рыбу выловили, разрубили и нашли в ее желудке бутылку из-под клико. Крепкую бутылку с трудом разбили и извлекли клочки бумаги, наполовину разъеденные морской водой.

Гленарван, рассмотрев внимательно бумажки, сказал: «Здесь три документа, по-видимому, копии одного и того же. Один написан по-английски, второй — по-французски, третий — по-немецки». Затем вся компания, как могла, занялась восстановлением текста, сопоставляя варианты на трех языках, и через некоторое время получился вот какой загадочный документ:

7 июня 1862 трехмачтовое судно «Британия» Глазго
потерпело крушение гони южн
берег два матроса
Капитан Гр дости
контин пл жесток инд
брошен этот документ долготы
и 37° 11 широты Окажите им помощь
погибнут

Затем необходимо было расшифровать записку, примысливая к обрывкам слов отсутствующие буквы. С подобными проблемами вы, наверное, сталкивались при заполнении кроссворда. Кое-что стало проясняться... Теперь приведу цитату — кульминацию завязки романа:

«Помолчав, Гленарван продолжал:

— Друзья мои, все эти предположения кажутся мне весьма правдоподобными. Полагаю, что катастрофа произошла у берегов Патагонии».

Из комплекта газет им стало известно:

— «Тридцатого мая тысяча восемьсот шестьдесят второго года. Перу. Кальяо. Место назначения Глазго, «Британия», капитан Грант».

— Грант! — воскликнул Гленарван. — Уж не тот ли это отважный шотландец, который мечтал основать новую Шотландию на одном из островов Тихого океана?

— Да, — ответил Джон Манглс, — это тот самый Грант. В тысяча восемьсот шестьдесят первом году он отплыл из Глазго на «Британии», и с тех пор о нем ничего не слышно.

— Нет никаких сомнений! — воскликнул Гленарван.— Это он! «Британия» отплыла из Кальяо тридцатого мая, а седьмого июня, спустя неделю, потерпела крушение у берегов Патагонии. Теперь вся история этой катастрофы раскрылась перед нами. Вы видите, друзья мои, мы нашли ключ к решению почти всей загадки, и единственным неизвестным теперь является долгота, где произошло крушение.

— Но долгота и не нужна! — заявил капитан Джон Манглс.— Зная страну и широту, я берусь найти место бедствия.

— Следовательно, нам все известно? — спросила леди Гленарван.

— Все, дорогая Элен, и я могу заполнить теперь то, что смыла морская вода, с такой легкостью, словно писал бы этот документ под диктовку самого капитана Гранта.

Тут Гленарван снова взял перо и, не колеблясь, написал следующее:

«*7 июня 1862 года трехмачтовое судно «Британия», из порта Глазго, затонуло у берегов Патагонии в Южном полушарии. Два матроса и капитан Грант попытаются достичь берега, где окажутся в плену у жестоких индейцев. Они бросили этот документ под ... градусами долготы и 37°11′ широты. Окажите им помощь, или они погибнут*».

— Хорошо! Хорошо, дорогой Эдуард! — воскликнула леди Элен.— Если этим несчастным суждено когда-нибудь вновь увидеть свою родину, то они будут обязаны вам своим спасением».

Итак, Гленарван выдвинул гипотезу о месте крушения «Британии» и, после знакомства с детьми капитана Гранта, принял решение — организовал путешествие в Патагонию. Цель этого путешествия — поиски пропавшей экспедиции, то есть, говоря на более холодном научном языке, проверка гипотезы о месте крушения, и в случае ее истинности, последующие действия для поисков экипажа «Британии». Теперь я должен пропустить 200 страниц приключений, от которых захватывает дух не только у двенадцатилетних, и подвести итог. Экспедиция капитана Гранта на берегах Патагонии обнаружена не была, и гипотеза Гленарвана оказалась ошибочной. Именно в это время случайный соучастник путешествия знаменитый географ Паганель предлагает иную трактовку некоторых из отрывков слов в записке капитана Гранта, и в результате: «Паганель, водя пальцем по отрывочным строкам документа, уверенно подчеркивая некоторые слова, прочел следующее:

— «*Седьмого июня тысяча восемьсот шестьдесят второго года трехмачтовое судно «Британия», из порта Глазго, потерпело крушение после ...*». Здесь можно вставить, если хотите: «*двух дней*», «*трех дней*» или «*долгой агонии*»,— все равно — это безразлично. «*... у берегов Австралии. Направляясь к берегу, два матроса и капитан Грант попытаются высадиться ...*», или «*высадились на континент, где они попадут ...*», или «*попали в плен к жестоким туземцам. Они бросили этот документ ...*» и так далее и так далее».

Выдвинута другая гипотеза, и руководитель экспедиции Гленарван принимает новое решение — «Дункан» держит путь к Австралии. Теперь вновь пропустим 400 страниц увлекательных и волнующих приключений. Однако и в Австралии капитан Грант не был обнаружен, и гипотеза Паганеля оказалась ошибочной. Все же бывший боцман «Британии» Айртон сообщил кое-какую новую информацию: капитан Грант незадолго до крушения собирался посетить Новую Зеландию. Еще до сообщения Айртона Паганель тоже понял ошибочность своей трактовки записки капитана Гранта и предложил новую:

— «*Двадцать седьмого июня тысяча восемьсот шестьдесят второго года трехмачтовое судно «Британия» из Глазго, после долгой агонии потерпело крушение в южных морях, у берегов Новой Зеландии* (по английски Zealand). *Двум матросам и капитану Гранту удалось добраться до берега. Здесь, терпя постоянно жестокие лишения, они бросили этот документ под... долготы и тридцать седьмым градусом и одиннадцатой минутой широты. Окажите им помощь, или они погибнут*».

Паганель умолк. Подобное толкование документа было вновь допустимо. Но именно потому, что оно было столь убедительным, как и первые толкования, оно могло быть столь же ошибочным.»

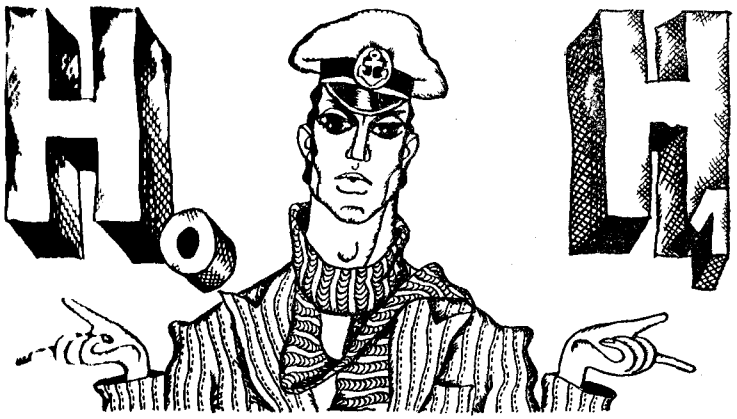
С последним нельзя было не согласиться — всякая гипотеза, даже весьма правдоподобная, пока не будет подтверждена, может оказаться ошибочной. И в самом деле, очередной вариант расшифровки текста записки, предложенный Паганелем, оказывается неверным вследствие ошибочности самой первой гипотезы Гленарвана. Напомню: он предположил, что тексты записки на разных языках — абсолютно одинаковы, а они были разными, ибо остров, на котором оказался капитан Грант, на английских и немецких картах значился как остров Марии-Терезы, а на французских — как Табор.

Итак, подведем итог действиям лорда Гленарвана. При получении информации Гленарван ее анализирует, выдвигает гипотезу и принимает решение о действиях, необходимых для ее проверки. В результате собирается новая информация, и она дает основания для принятия или отклонения гипотезы, ибо она может оказаться верной или ошибочной. Затем, в зависимости от результатов проверки гипотезы и поступившей новой информации, Гленарван вновь анализирует ситуацию, выдвигает новые гипотезы, принимает решения и т. д. Все это продолжается вплоть до достижения цели или признания невозможности ее достичь и принятия решения о прекращении дела. Это и есть процесс управления. Управление кораблем, заводом, установкой первичной переработки нефти, школой или любым другим объектом происходит аналогично. Сначала выдвигаются гипотезы (корабль идет по курсу или отклонился от него), затем, на основании имеющейся информации, гипотезы проверяются (производятся наблюдения, измерения технологических параметров) и затем принимаются решения о необходимых действиях (так держать или лево руля и т. д.).

При коронации английской королевы, церковной службе или исполнении оперы «Евгений Онегин» никакой неопределенности нет: ход действия, все слова и поступки заранее предопределены, и управление, по существу, отсутствует. Конечно, в момент, когда страдающая Татьяна в последнем акте оперы опускается в кресло, дабы дать возможность Онегину преклонить колено, может подняться гневный вой — Татьяна вошла в роль полностью и села на уютно устроившуюся в кресле тощую закулисную кошку — и ход действия будет несколько нарушен. Но все же Татьяна, несмотря на смех в зрительном зале, не улыбнется и не бросится в объятия к Онегину, а, прогнав кошку, будет продолжать страдать, и Онегин, прежде чем за кулисами обсуждать накладку, споет «О жалкий жребий мой...».

Подобные ситуации все же еще раз подчеркивают, что управление предполагает наличие неопределенности и возможность выбора.

Поэтому, говоря об управлении, мы будем подразумевать наличие неопределенности, в условиях которой происходит последовательный процесс выдвижения и проверки гипотез, принятия решений и проверки их правильности. И нам предстоит обсудить некоторые аспекты этого, порой достаточно сложного процесса.



• • • ▶ Остап Бендер принимает решение

Вспомните завязку знаменитого «Золотого теленка» Ильи Ильфа и Евгения Петрова:

«На безмятежном невспаханном лбу Балаганова обозначилась глубокая морщина. Он неуверенно посмотрел на Остапа и промолвил:

— Я знаю такого миллионера. Может выйти дело.

С лица Бендера мигом сошло все оживление. Лицо его сразу же затвердело и снова приняло медальные очертания.

Идите, идите,— сказал он,— я подаю только по субботам, нечего тут заливать.

Честное слово, мосье Бендер...

Слушайте, Шура, если уж вы окончательно перешли на французский язык, то называйте меня не мосье, а ситуайен, что значит — гражданин. Кстати, адрес вашего миллионера?

Он живет в Черноморске.

Ну конечно, так и знал. Черноморск! Там даже в довоенное время человек с десятью тысячами назывался миллионером. А теперь... могу себе представить! Нет, это чепуха!

Да нет же, дайте мне сказать. Это настоящий миллионер. Понимаете, Бендер, случилось мне недавно сидеть в тамошнем допре...»

После этого доверительного разговора возникли две гипотезы: *в Черноморске есть подпольный миллионер Корейко* и ей противоположная — *никакого миллионера нет*, а Балаганов, мягко говоря, ошибается.

Гипотезы обычно обозначают буквой H (от греческого Hypotesis). Исходную гипотезу или нуль-гипотезу *Корейко — подпольный миллионер* будем обозначать H_0, а конкурирующую или альтернативную, то есть противоположную исходной, *Корейко не миллионер, а счетовод второго разряда* будем обозначать H_1. Остап — человек действия, и он должен был принять решение — сказать либо «ДА», либо «НЕТ», то есть отдать предпочтение одной из двух гипотез, и, приняв гипотезу, действовать дальше так, как будто принятая гипотеза верна.

Если Остап говорил «ДА», то есть принимал гипотезу H_0, надо было действовать — устремляться в Черноморск и использовать для изъятия миллиона один из ему известных четырехсот способов сравнительно честного отъема денег. Если же он говорил «НЕТ» и принимал альтернативную гипотезу H_1, то не нужно было действовать.

Однако, какое бы Остап ни выбрал решение, он мог ошибиться. Разберем возможные ошибки Остапа. Ошибка, при которой нуль-гипотеза на самом деле верна (*Корейко — миллионер*), а принимается альтернативная гипотеза (*Корейко — простой счетовод*), называют ошибкой первого рода или пропуском (здесь Остап «пропускает» миллионера, за которым охотится). Эту ошибку Остап совершит, если он откажется от «охоты» на Корейко, не поверив Балаганову, в ситуации, когда Корейко на самом деле подпольный миллионер.

Но, может быть, Корейко уже не миллионер (на жизненном пути миллионеров, и особенно подпольных, столько препятствий и непредвиденных катастроф!), или, возможно, Шура Балаганов ошибся, и на самом деле Корейко — счетовод. Если при этом Остап принимает гипотезу о наличии у Корейко миллионов, то он потратит время и весьма скудные наличные средства на поиски несуществующего миллионера и в результате лишь познакомится со счетоводом, с которого не получишь даже «возмещения расходов». Это и есть ошибка второго рода, или ложная тревога. Великого комбинатора, рвущегося к «блюдечку с голубой каемочкой», на котором должны принести ему миллионы, такой исход, конечно, не устраивает — жаль времени.

Придирчивый читатель, небось, скажет, что можно было бы принять за нуль-гипотезу как раз отсутствие миллионов у Корейко, а за альтернативную Корейко — миллионер. Конечно, можно. Однако Остапу очень хочется миллион, а миллионы не так уж часто попадаются на дороге, и ему важнее избежать потери миллионов, чем потратить зря время на прогулку в Черноморск — все равно в данный момент делать нечего. Обычно называют ошибкой первого рода ту, которую важнее избежать. Если же ошибки имеют одинаковую значимость, а это редко встречается, безразлично какую из них принять за ошибку первого рода. Сведем в таблицу все возможные варианты.

Таблица 1

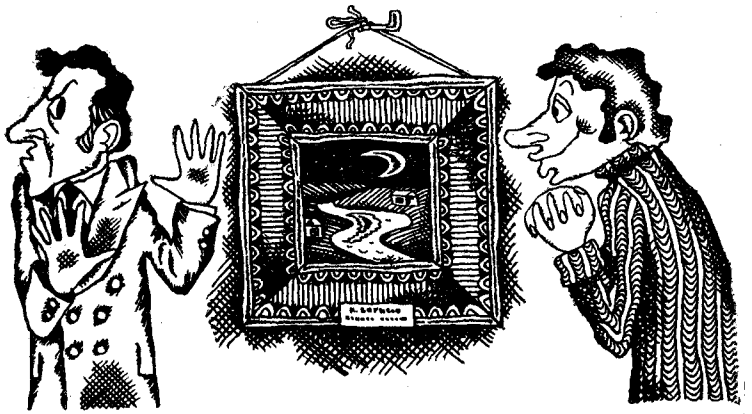
Решение Остапа	Истинное положение	
	Подпольный миллионер	Простой счетовод
Подпольный миллионер	Верно	Ошибка 2-го рода — ложная тревога
Простой счетовод	Ошибка 1-го рода — пропуск миллионера	Верно

Итак, для Остапа ошибки первого и второго рода совершенно не равноценны. Поэтому Остап выбирает гипотезу H_0 и организует знаменитый поход.

Я напомнил вам об Остапе Бендере для демонстрации встречающейся на каждом шагу ситуации: на основании каких-то соображений или данных выдвигаются несколько гипотез, и наблюдателю или руководителю, исследователю, или участнику игры нужно принять решение — отдать предпочтение одной из этих гипотез и действовать далее в соответствии с принятым решением. Напомню: при выходе из дому встретить первой блондинку — событие, причем статистически устойчивое. Можно выдвинуть такие гипотезы: *вероятность встретить первой блондинку меньше* 0,1; *брюнетки встречаются чаще, чем блондинки; вероятность встретить трех блондинок подряд больше* 0,01. Подобные гипотезы называют статистическими, ибо здесь события обладают статистической устойчивостью и можно осуществить их проверку статистическими методами.

В то же время гипотезы о наличии или отсутствии миллионов у Корейко — гипотезы не статистические, их нельзя проверить статистическими методами, ибо событие *Корейко — подпольный миллионер* не случайное, а неопределенное. И при принятии решения Остап опирается не на статистические данные, а лишь на кое-какую информацию нестатистического характера и интуицию. Но мне предстоит рассказать о том, как при проверке гипотез для принятия возможно более обоснованных решений, опирающихся не на интуицию, а на данные наблюдений или экспериментов, используются методы математической статистики.

Вскоре мы обратимся к этим проблемам.



Немного о критериях

Великолепную экспозицию живописи из личной коллекции Хаммера мне удалось посмотреть в Одессе летом семьдесят третьего года. На выставке я был несколько раз, и, кроме эмоционального наслаждения, получил еще одно сильное впечатление. Экспансивные и остроумные одесситы с завидной откровенностью обсуждали полотна великих мастеров, и их хлесткие, едкие и иногда смешные комментарии были весьма разнообразны. Одинаковых мнений не встречалось, и если один приходил в восторг от «Сеятеля» Ван-Гога, то другой с возмущением осуждал фиолетовую гамму цветов и говорил, что лично у себя он эту картину не повесил бы даже по прямому указанию начальника милиции.

Таким образом, оценки оказались крайне субъективными, и совершенно невозможно было бы установить, каковы критерии этих оценок: одни говорили о вялости рисунка, другие — о гениальной насыщенности красок, третьи — о перенасыщенности — на всех не угодишь...

Когда на школьном дворе между мальчишками классов 6-А и 6-Б — московские школьники говорят «ашки» и «бешки» — разгорается острый спор на тему «Какой класс выше ростом?», то трудно принять чью-нибудь сторону. В шестом классе, конечно, девчонки в счет не идут, и спор идет лишь о росте мальчишек.

Ашки кричат, что их Алеха выше всех бешек. Но бешки гудят, что это вовсе ничего не означает, подумаешь, в классе один — верзила... Зато почти все остальные бешки

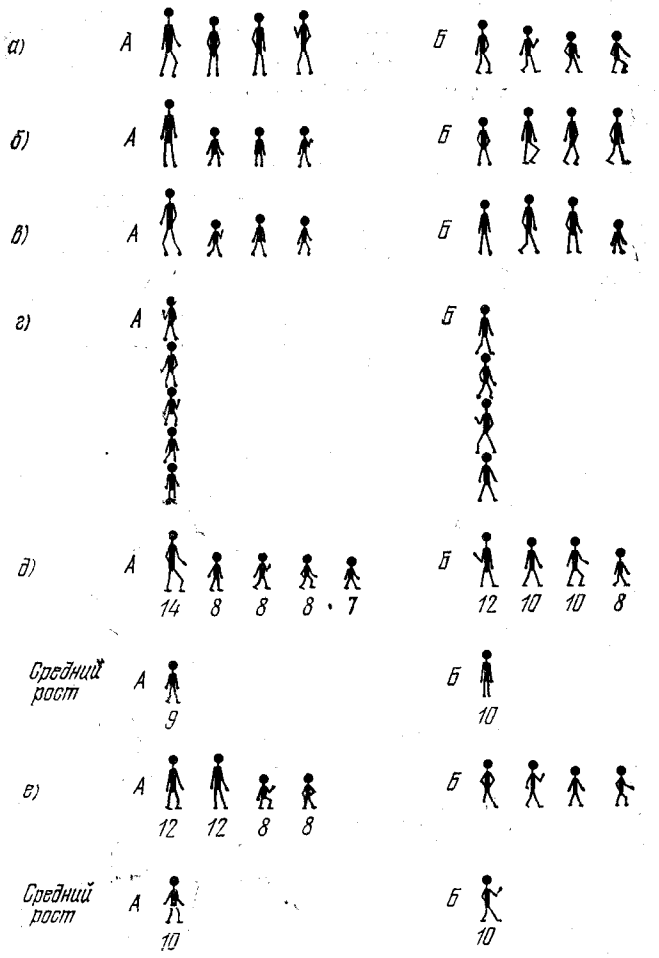


Рис. 1.

выше ашек! Обычно такой спор приводит к потасовке, и причиной тому служит полная неясность предмета спора.

В самом деле, как следует понимать высоту класса? Какой смысл вкладывают ребята в эту величину, как ее

определить? Так как обычно никто толком ответить на эти вопросы не может, то я предложу на выбор несколько вариантов. Вот они.

Ашки считаются выше бешек, если:

а) *Любой из ашек выше любого из бешек* (рис. 1, *а*). Если именно так обстоит дело, то ситуация очевидная — спорить не о чем. Но едва ли так обстоит дело в действительности.

б) *Самый высокий из ашек выше самого высокого из бешек* (рис. 1,*б*). Здесь может оказаться, что один из ашек высок ростом, но все остальные — маленькие, например все остальные ашки ниже любого из бешек. Такой вариант решения спора следует, пожалуй, отвергнуть, хотя при баскетбольном матче между классами именно самый высокий может сыграть решающую роль.

в) *Для любого из ашек найдется ученик из класса Б ниже его ростом* (рис. 1,*в*). В такой обстановке может оказаться среди бешек один совсем маленький, и хотя остальные бешки рослые, и даже превосходят всех ашек, кроме верзилы, но ашки все же впереди. Этот вариант определения тоже не очень-то отвечает интуитивному понятию высоты группы людей.

г) *Сумма ростов ашек больше суммы ростов бешек.* На рис. 1, *г* ашки оказались выше. Такая ситуация может иметь место по разным причинам, например за счет того, что каждый из ашек выше, но, главное, такое положение может оказаться вследствие большего количества мальчиков в классе А, чем в классе Б. Возможно, при определении роста такой критерий и не кажется отвечающим существу дела, но если спор зашел бы о силе классов при игре в перетягивание каната, то критерий — суммарная сила — оказался бы вполне удовлетворительным.

д) *Средний рост класса А больше среднего роста класса Б.* Здесь, как видите, вычисляются средние арифметические величины ростов для каждого класса и затем они сравниваются (рис. 1,*д*). Теперь все определяет соотношение ростов в пределах каждого класса, и ашки проигрывают. Верзила не спасает положение, ибо остальные ашки в общем-то маленькие, в то время как бешки примерно одинаковые и чуть выше основной массы ашек.

е) Как видите, спор действительно решить не просто, нужны еще аргументы для выбора одного из критериев. Особенно усложняется обстановка, если будет такая картина, как на рис. 1, *е*. Средний рост у обеих групп одинаков, но бешки примерно одинаковые, в то время как ашки имеют рослых и маленьких, и они уравновешивают друг друга.

Мурат шкрей ФЗ

Если вновь просмотреть все варианты, то все же средний рост лучше других характеристик отвечает нашему интуитивному пониманию роста группы, и в споре мальчишек, пожалуй, естественнее всего принять за критерий именно его.

Рост школьника шестого класса, конечно, следует рассматривать как случайную величину, и поэтому средний рост класса — это среднее выборки, или эмпирическое среднее. Оно служит оценкой математического ожидания самой случайной величины, то есть роста школьника. Как мы договорились с самого начала, понятие математического ожидания считается известным, и я о нем упомянул сейчас для некоторого его обобщения.

Среди школьников шестого класса есть блондины, брюнеты, шатены, рыжие, и можно рассматривать распределение вероятностей другой случайной величины — роста рыжих школьников шестого класса. Это уже будет условное распределение: распределение вероятностей роста школьника шестого класса при условии, что школьник рыжий. Вообще говоря, условное распределение и исходное — безусловное распределение будут различаться. Каждое из этих распределений имеет свое математическое ожидание: величины среднего роста шестиклассников и рыжих шестиклассников могут не совпадать.

Величина среднего роста рыжих шестиклассников — условное математическое ожидание — это также критерий, если сравнивать, скажем, рост рыжих шестиклассников из разных школ.

Условное математическое ожидание — важное понятие, и мне в дальнейшем придется им воспользоваться.

Когда учитель в шестом классе выставляет оценки за диктант, то он должен подсчитать общее число ошибок и выставить пять, если ошибок нет, или два, если их достаточно много,— здесь как будто критерий вполне четко определен инструкцией Министерства просвещения. Однако учитель сознательно или подсознательно различает не только грубые и негрубые ошибки, но и не очень грубые, описки и пропуски букв, а подчас сравнительные продвижения или падения подростка, и в результате «нарушает» правила: то ставит пять с минусом, то три с двумя минусами вместо двойки. Я не буду осуждать подобные поступ-

ки учителя: мой собственный более чем тридцатипятилетний педагогический опыт дает возможность утверждать, что неформальный по существу педагогический процесс плохо формализуем, а всякие бюрократические инструкции — не подмога, а скорее помеха для квалифицированного преподавателя, а неквалифицированному нужны не инструкции типа «Правил дорожного движения», а помощь в повышении квалификации и педагогического мастерства.

Итак, при классификации знаний школьников на четыре условных градации (или пять, если допустить «кол») происходит проверка гипотезы об отнесении данного диктанта или данного ученика к одной из этих градаций. Как ни плох критерий количества ошибок, все же он есть и выражается числом.

Однако подчас выставляется две отметки — одна за количество орфографических ошибок, другая — за количество пунктуационных, или даже три — за сочинение, когда еще добавляется оценка за содержание. Какой вы предложите критерий для сравнения сочинений и выбора среди них лучшего? Как, например, выбрать лучшую из работ при таких их оценках: 2/1/4; 0/1/3; 4/0/5. Не знаю, как вы, но я находился бы в весьма затруднительном положении при необходимости осуществить выбор.

Именно, вследствие неоднозначности возможных решений критерий должен выражаться одним числом.

Попробуем выдумать критерий, предоставляющий нам возможность на основании имеющихся данных однозначно решать вопрос о выборе лучшего из этих сочинений.

Замечу: чем больше два первых числа — число орфографических и число синтаксических ошибок,— тем хуже сочинение, но чем выше оценка за содержание,— тем лучше. Здесь в качестве критерия можно выбрать отношение

$$R = \frac{\text{отметка за содержание}}{\text{орфогр.} + \text{синтак.} + 1},$$

где в знаменателе помещено суммарное число ошибок, а единица здесь приписана, чтобы не обратить в бесконечность величину R, когда ошибок вовсе нет.

Тогда для приведенных трех примеров имеем следующие значения критерия R, соответственно:

$$R_1 = \frac{4}{2+1+1} = 1, \quad R_2 = \frac{3}{0+1+1} = \frac{3}{2}, \quad R_3 = \frac{5}{4+0+1} = 1.$$

Таким образом, по критерию R оказывается лучшей вторая работа, а первая и третья одинаковы.

Однако мне, скажем, больше импонирует последняя работа: все же пятерка за содержание — это заметно лучше, чем тройка, даже при немного большем числе ошибок. Легко предложить другой критерий, дающий преимущество оценке за содержание. Для этого в знаменатель добавим не единицу, а пять, и получим

$$K = \frac{\text{отметка за содержание}}{\text{орфогр.} + \text{синтак.} + 5}.$$

Теперь для тех же примеров значения критерия K дают

$$K_1 = \frac{4}{2+1+5} = \frac{1}{2}, \quad K_2 = \frac{3}{0+1+5} = \frac{1}{2},$$

$$K_3 = \frac{5}{4+0+5} = \frac{5}{9},$$

то есть по критерию K лучшей оказывается уже последняя работа, а первая и вторая — одинаковые.

Думаю, теперь вы сможете без труда предложить критерии, по которым лучшей будет первая работа или все будут равнозначны — словом, можно получить, так сказать, что угодно для души.

Итак, в одной и той же задаче при выборе лучшего объекта или лучшего решения всегда можно предложить множество самых различных критериев. При этом выпукло видно, как существенно судьба предмета обсуждения или спора, конкурсных работ, а то и человека или коллектива зависит от выбранного критерия и сколь эфемерны, подчас, рассуждения о справедливых и несправедливых решениях, когда оценка ситуации и выбор критерия достаточно произвольны.

Не пропустить бы радиозайчик

Когда кромешной ночью вы добираетесь домой через лес или стройку, а батарейка карманного фонарика уже на исходе, то вы периодически включаете фонарик на короткое время — нужно лишь успеть увидеть: есть что-то впереди или путь свободен, дабы не расшибить себе лоб или не сломать ногу. Если впереди что-то есть, луч света от фонарика отразится, и вы увидите «зайчик» — обнаружите препятствие, хотя и не сможете сказать определенно, дерево ли это, бетонная плита или корова. Если же препятствия нет, зайчик уйдет в черную пустоту, и можно сделать несколько шагов вперед. Впрочем, вы можете и ошибиться: то в страхе остановитесь, когда на самом-то деле мелькнет свет от далекого окошка, то впереди окажется наклоненный лист железа, и зайчик отразится куда-то вбок, так что вы, чертыхаясь, ушибетесь о проклятую деталь.

Итак, всякий раз имеется две гипотезы: H_0 — *препятствия нет,* и H_1 — *имеется препятствие.* При этом можно совершить ошибки первого и второго рода — остановиться, опасаясь препятствия, когда его на самом деле нет, и не заметить препятствие, когда оно действительно есть.

Разберем теперь значительно более серьезную проблему принятия решения в радиолокации. Первейшая задача наземной радиолокационной станции — обнаружение самолета, когда он появляется в пределах видимости. Конечно, у радиолокации есть и другие задачи: определение координат самолета, его скорости и ряд других — они зависят от назначения станции. Но я буду сейчас обсуждать только проблему обнаружения. На кораблях также установлены радиолокаторы, и они должны обнаруживать другой корабль или айсберг, обнаруживать своевременно, чтобы успеть принять меры и избежать столкновения.

Если не вдаваться в технические детали, то идея радиолокации — это запуск того же зайчика, только вместо света используются короткие посылки электромагнитных излучений другого диапазона с большей длиной волны, чем у света. А приемная антенна радиолокатора — аналог глаза: именно она принимает отраженный от препятствия радиозайчик. Принятый сигнал имеет очень маленькую мощность и просто так его не обнаружишь — в приемной аппаратуре необходимо обеспечить большое усиление. Но мешают собственные шумы аппаратуры, атмосферные разряды, разные технические помехи. Собственные шумы есть в любой приемной аппаратуре. Скажем, когда вы слушаете передачу по обычному радиоприемнику, то, когда замолкает диктор, слышно шипение — это и есть всегда присутствующий фон. Все это вместе приводит к ошибкам.

Но прежде чем о них пойдет речь, хочу обратить ваше внимание на задачу обнаружения в системе противовоздушной обороны, еще более ответственную, чем в системе гражданской аэродромной службы. С помощью комплекса технических средств в систему ПВО поступает периодически информация о состоянии охраняемого пространства: появились ли угрожающие объекты, например самолеты или ракеты противника, или их нет.

Информация достается не даром: служба обнаружения весьма сложная и дорогостоящая. И при этом она тоже не безошибочна: сигналы в системе обнаружения ПВО могут быть приняты за шум, который, как я уже заметил, всегда есть и на фоне которого сигнал следует обнаружить. В таком случае имеет место ошибка второго рода — пропуск жизненно важной информации, жизненно важной не в переносном, а, как вы понимаете, в самом прямом смысле, ибо при космических скоростях противовоздушной обороне следует принимать меры немедленно.

Ложная тревога — ошибочное заключение об обнаружении вражеского объекта, когда его на самом деле нет,— также не безопасна: это же может означать команду противовоздушным силам принять необходимые меры защи-

ты, а они, конечно, предусматривают и ответные меры подавления средств нападения противника. Нет, уж лучше постараться избежать такой тревоги...

Система радиолокационного обнаружения может быть построена различным образом не только в смысле технической реализации, но и по методам обработки данных, и качественные характеристики разных систем будут различными. Чем же характеризовать качество работы системы радиолокационного обнаружения?

Давайте разберемся. При каждом цикле работы радиолокатора посланный передающей антенной импульс может отразиться от объекта и быть принят или пропущен; в отсутствие объекта импульс уйдет в пространство, а приемная станция может зафиксировать его отсутствие или принять шум за сигнал. Важно сейчас отметить, что шум — это процесс случайный, и вся работа системы радиолокационного обнаружения происходит в условиях статистической устойчивости. Поэтому гипотезы — H_0 *есть только шум* и H_1 *есть и сигнал, и шум* — это статистические гипотезы, и имеет смысл говорить о вероятностях ошибок первого и второго рода. В наших условиях ошибка первого рода — это ситуация, когда никакого объекта нет, иначе говоря, есть только шум, а принимается гипотеза H_1 о наличии сигнала и шума, то есть ложная тревога. В математической статистике вероятность ошибки первого рода или вероятность ложной тревоги носит название уровня значимости. Такой термин более удачен: значимость принятия ошибочного решения о наличии объекта, когда его нет, понятие довольно ясное, а здесь оно лишь усовершенствуется — ему придается понятная численная мера.

Теперь всю ситуацию можно представить в виде таблицы 2, аналогичной таблице 1.

Однако следует подчеркнуть существенную разницу между первой и второй таблицами: хотя Остап Бендер может оценить потери при ошибках первого и второго рода, но никакой вероятности этих ошибок нет, ибо *Корейко — миллионер* — событие неопределенное, а не случайное, да и поступки Остапа, пусть он и действует по воле случая, не имеют никакого распределения вероятностей.

Итак, задача системы радиолокационного обнаружения состоит в том, чтобы отдать предпочтение одной из двух гипотез, принять одно из двух возможных решений: *есть только шум* или *есть сигнал и шум*, сказать «НЕТ» или «ДА».

Можно предложить множество методов принятия решения в подобной ситуации. И можно, например, считать лучшим из двух таких методов тот, который обеспечивает меньший уровень значимости, то есть меньшую вероятность ошибки первого рода.

Таблица 2

Принятое решение	Истинное положение	
	Шум	Сигнал + шум
Шум	Верно. Вероятность правильного решения 1-α	Ошибка 2-го рода — пропуск сигнала. Вероятность ошибки β
Сигнал + шум	Ошибка 1-го рода — ложная тревога. Вероятность ошибки — уровень значимости α	Верно. Вероятность правильного решения 1-β

Теперь конечно, следует обсудить вопрос о допустимых или возможных величинах уровня значимости. Но прежде чем приступить к этому важному вопросу, мне хочется указать на другие проблемы, связанные со статистической проверкой гипотез.

Вы, конечно, слышали об азбуке Морзе, где буквы кодируются различными комбинациями точек и тире. Их можно заменить любыми двумя заметно различающимися сигналами. В современной телеграфии применяются либо посылки тока и паузы, либо посылки постоянного тока разной полярности, либо импульсы переменного тока одинаковой длительности, но разной частоты или фазы. Для всех подобных методов кодирования главное — это наличие двух различимых сигналов. Их можно, как это принято в теории электронных вычислительных машин, записывать как 0 и 1. Тогда каждая буква — это комбинация символов — нулей и единиц. Скажем, из пяти символов можно составить $2^5=32$ различных комбинаций, то есть 32 буквы.

Если символы передаются по электрическому каналу связи, то при любом из указанных методов в результате действия помех, присутствующих в любом канале связи, нуль может быть расшифрован как единица или едини-

ца — как нуль. Вся картина формально та же, что и при радиолокационном обнаружении: если H_0 — это передача нуля, а H_1 — передача единицы, то ошибка первого рода — это принятие единицы, когда был передан нуль, а ошибка второго рода — принятие нуля, когда передавалась единица. Здесь уровень значимости — это вероятность принять единицу, когда передавался нуль.

Аналогичная ситуация складывается при контроле качества продукции. Показатели качества продукции весьма разнообразны. Так, для ботинок, автопокрышек, электрических лампочек главнейший показатель — срок службы!

И здесь складывается парадоксальная ситуация: с одной стороны, не стоит покупать ботинки, если они придут в негодность через неделю, но, с другой стороны, нельзя проверить срок службы ботинок, не сносив до дыр, а тогда их не только не обменяешь, но и не продашь.

Срок службы подобной продукции можно установить лишь по аналогии. Скажем, из недельной или месячной партии продукции делают выборку, то есть отбирают несколько автопокрышек по определенному правилу. Выборка подвергается проверке на срок службы. Например, отобранные автопокрышки ставят на контрольные автомашины и проверяют, сколько километров они могут пройти до того, как протектор станет «лысым». Если все или почти все покрышки из выборки прошли пятьдесят тысяч километров, то имеются некоторые основания считать, что остальные автопокрышки из той же партии также пройдут не менее пятидесяти тысяч. Иначе говоря, можно полагать качественные показатели изделий всей партии близкими к показателям выборки, подвергшейся контролю.

Итак, контроль качественных показателей в случае, когда после контрольных испытаний продукция гибнет, как, например, снаряды, или уже оказывается негодной, как ботинки или автопокрышки, всегда организован как выборочный. Впрочем, выборочный контроль разумно принять и в ситуации, когда сам контроль не приводит к порче продукции. Не будете же вы настаивать на проверке размеров каждой спички или каждого гвоздя, ибо такой контроль, очевидно, не рентабелен: в подобной ситуации стоимость контроля может превысить стоимость самой продукции. Поэтому выборочный контроль качества продукции широко распространен.

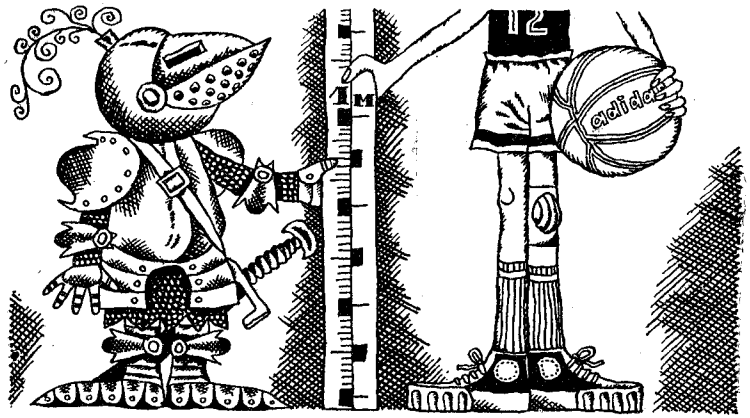
При приемочном контроле продукции имеются две альтернативные гипотезы: либо *партия изделий годная* (H_0), либо *негодная, бракованная* (H_1). Задача контроля — отдать предпочтение одной из этих гипотез и, естественно, принять одно из двух возможных решений: пропустить или забраковать партию — сказать «ДА» или «НЕТ». Как видите, полная аналогия с задачей радиолокационного обнаружения.

Мне хочется подчеркнуть: всякий подобный метод контроля — это метод статистической проверки гипотез, то есть метод принятия решений в условиях неопределенности, и притом неопределенности, характеризующейся статистической устойчивостью. Такой подход приводит к далеко идущей трактовке методов математической статистики. В литературе последних десятилетий встречается даже такое определение: содержанием математической статистики является разработка методов принятия решений в условиях неопределенности, характеризующейся статистической устойчивостью. Конечно, этого уже было бы достаточно для отнесения математической статистики к перворазрядным разделам науки. Впрочем, мне представляется содержание математической статистики более глубоким, но не будем сейчас тратить время на общие рассуждения.

Требования к качеству продукции существенно зависят от ее назначения: если при производстве спичек или гвоздей можно допустить, скажем, брак в 5%, то при производстве самолетных двигателей или лекарств качество должно быть значительно лучше. А следовательно, различными должны быть и требования к методам контроля.

Как я уже говорил, любую из альтернативных гипотез можно принять за нуль-гипотезу, и это решение определяется позицией самого исследователя.

Будем сейчас оценивать ситуацию с позиций поставщика, для которого браковка годной партии — наименее приятный факт, и примем за ошибку первого рода браковку годной продукции. Когда вероятность ошибки — уровень значимости равен α, это означает, что в длинной серии проверок в $100\alpha\%$ случаев будет отвергнута правильная гипотеза. Так, если $\alpha=0{,}02$, то $100 \cdot 0{,}02=2$, и в среднем в двух процентах случаев будет отвергнута правильная гипотеза. Мерой доверия к утверждению *верна гипотеза* H_0 (в данном случае H_0 — *партия годная*) служит вероятность $1-\alpha$ — доверительная вероятность.



Отношение правдоподобия

Займемся практической и важной для нас сегодня проблемой акселерации.

Акселерация (от латинского слова acceleratio — ускорение) — это ускоренное развитие: увеличение роста — длины тела, как говорят антропологи, определенных этнических или профессиональных групп населения и более раннее половое созревание.

Вы, наверное, слышали об акселерации. Если четких данных вы не имеете, но хотите проверить достоверность фактов, то можете провести исследование самостоятельно.

Многолетние наблюдения показывают, что характер распределения роста людей не меняется от поколения к поколению — распределение это нормальное. Поэтому естественно выяснить, меняются ли числовые характеристики распределения. Обсудим здесь лишь среднее значение. Итак, надо выяснить, меняется ли от поколения к поколению величина среднего роста, считая распределение роста нормальным.

На основании достаточно достоверных статистических данных относительно поколения родившихся в 1908—1913 годах можно определить величину среднего роста взрослых мужчин (в возрасте двадцать — тридцать лет): она равна 162 см. Сопоставим это число со средним ростом мужчин другого поколения. Скажем, у вас имеется возможность измерить или узнать рост нескольких десятков мужчин, ро-

дившихся в течение пятилетнего периода между 1943 и 1948 годами. После подсчета (который, конечно, каждый может без труда произвести), средний рост этой группы лиц оказывается равным 170 см. Гипотеза, подлежащая проверке, состоит в том, что средний рост от поколения к поколению не меняется, а имеющиеся колебания (такие, как 162 см и 170 см) — это «происки дьявола», то есть просто результат естественного разброса, поскольку средний рост — случайная величина. Примем в качестве нульгипотезы *средний рост в поколении родившихся в 1943 — 1948 годах тот же, что и в поколении родившихся в 1908 — 1913 годах;* при альтернативной гипотезе — *средний рост за эти тридцать пять лет увеличился.*

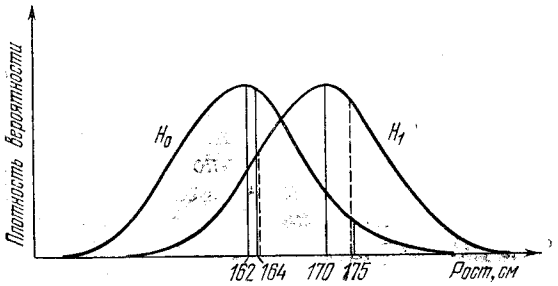


Рис. 2.

Мне хочется продемонстрировать метод рассуждений и, дабы не слишком усложнять ситуацию, я откажусь от столь общей формулировки альтернативной гипотезы и приму ее более простой: *средний рост поколения 1943 — 1948 годов равен 170 см,* и, следовательно, за 35 лет увеличился на 8 см. На рис. 2 представлены плотности вероятностей роста при каждой из гипотез, причем «левая» кривая соответствует исходной гипотезе.

Итак, обратимся к группе наблюдаемых мужчин более позднего поколения и проверим рост первого из них (скажем, первого по алфавиту) — пусть он оказался 164 см. Такой рост может быть у представителя любой из двух групп. Рассмотрим обе возможности. Если он из «левой» группы, то он обладает ростом 164 см с плотностью вероятности, равной высоте отрезочка, параллельного вертикальной оси, над числом 164 до левой кривой, то есть более высокого отрезочка. На рис. 2 он вычерчен сплошной ли-

нией. Если же наблюдаемый мужчина принадлежит к группе с предполагаемым новым распределением вероятностей, график плотности которого — «правая» кривая, то он будет обладать ростом 164 см с плотностью вероятности, равной высоте аналогичного отрезочка над числом 164, но уже до правой кривой — он отмечен пунктиром. Нам надо принять решение — отнести наблюдаемое значение к одному из этих двух распределений. Предлагаемый принцип принятия решения заключается в выборе того из распределений, к которому наблюдение относится с большей вероятностью. В данном случае больше вероятность для нуль-гипотезы H_0 (больше высота столбика), и, если бы других наблюдений не было, нам следовало бы принять гипотезу H_0 о неизменности распределения ростов.

Читатель может обратить внимание на подмену вероятностей плотностью вероятности, но это законная подмена: при сравнении вероятностей удобно взять их отношение и сравнивать его с единицей, и тогда законно заменять вероятности на соответствующие плотности. Отношение плотностей вероятности носит название *отношения правдоподобия*. Название, как вы видите, вполне отвечает существу дела: отношение правдоподобия сравнивается с единицей, и в зависимости от того, оказывается ли оно больше или меньше единицы, выбирается «левая» или «правая» гипотеза, то есть выбирается более правдоподобное решение.

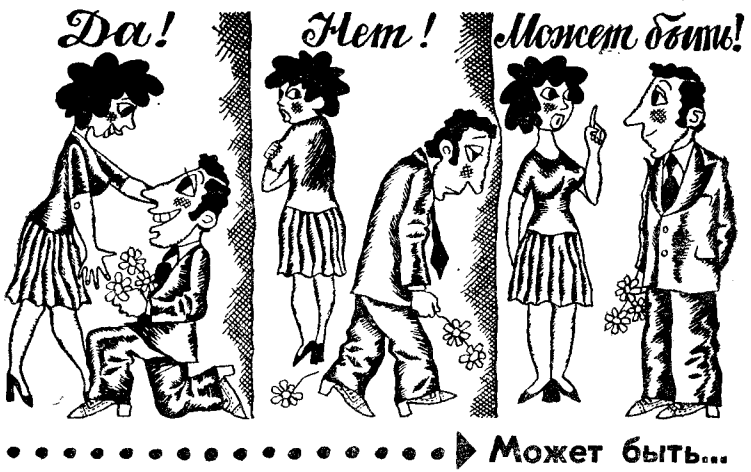
Теперь, конечно, следует указать на сложности, стоящие на нашем пути. Если рост второго индивидума из наблюдаемой группы 175 см, то с помощью тех же рассуждений следует принять гипотезу о его принадлежности к «правому» распределению — здесь высота столбика до «правой» кривой, соответствующей гипотезе H_1, больше, чем до «левой» кривой, и, таким образом, более правдоподобной представляется гипотеза H_1 об увеличении роста.

Сопоставление результатов двух проведенных наблюдений нас ставит в затруднение: какому же отдать предпочтение? Но более разумный подход состоит в совместном использовании обоих наблюдений или всех вместе, если их несколько. Для этого «точечные» отношения правдоподобия перемножают и получают выражение, которое также называется отношением правдоподобия, но уже для всей совокупности наблюдений. Теперь назначается число, называемое порогом, и правило принятия или отвержения гипотезы H_0 состоит в сравнении полученного отношения

правдоподобия с порогом: если оно больше порога, принимается гипотиза H_0 («левая» кривая), в противном случае принимается гипотеза H_1 («правая» кривая). В частности, в проблеме акцелерации данные очень убедительные, и действительно, средний рост поколения 1943—1948 гг. на 8 см больше среднего роста поколения 1908 — 1913 годов.

В обсуждаемой проблеме в действительности нет необходимости прибегать к критерию отношения правдоподобия: здесь очень обширный статистический материал, и можно обойтись более простыми средствами. Но есть задачи, где применение этого критерия весьма целесообразно и дает хорошие результаты, например при радиолокационном обнаружении — задаче, которой мы уже уделили много внимания, критерий отношения правдоподобия применяется с успехом.

Шумы, как это хорошо известно экспериментаторам, подчиняются нормальному закону распределения с нулевым средним значением, а сигнал вместе с шумом подчиняется тому же распределению, но с другим средним значением — оно равно как раз величине амплитуды импульса сигнала. Вот здесь и применяется основанная на критерии отношения правдоподобия специальная обработка поступившего сигнала для принятия столь важного решения: есть сигнал и шум или сигнала нет, а есть только шум. Аналогично в технике связи или управления, когда нужно обнаружить сигнал или, говоря другим языком, проверить нуль-гипотезу о наличии только шума при альтернативной гипотезе о наличии сигнала и шума вместе, широко и успешно применяется критерий отношения правдоподобия.



Вернемся к проблеме сравнения методов статистической проверки гипотез. Логика нашего обсуждения, как вы, по-видимому, догадались, приводит к тому, чтобы использовать уровень значимости как критерий качества методов проверки: чем меньше уровень значимости, тем лучше метод проверки. Однако в наших рассуждениях незаслуженно забыта ошибка второго рода, которая в задаче браковки представляет собой приемку бракованной партии, а в задаче радиолокационного обнаружения — пропуск сигнала. Ее вероятность мы обозначили β. Тогда мерой доверия к утверждению *верна гипотеза* H_1 служит вероятность $1-\beta$.

Дело в том, что при выбранном методе проверки гипотез вероятности ошибок первого и второго рода оказываются зависимыми, и их нельзя задать произвольно. Я поясню это утверждение на примере телеграфной связи, когда в качестве символов служат посылка тока и пауза. На рис. 3 показаны различные ситуации. Шум искажает сигнал, и поступающий в приемную аппаратуру суммарный сигнал выглядит так, что не ясно, посылка ли это тока или пауза. Простейший способ обработки такого сигнала в приемной аппаратуре — установление некоторого порога: если принятый сигнал превзошел порог, принимается решение о наличии посылки тока, если он оказался ниже порога, принимается решение о наличии паузы. При шуме

определенной интенсивности вероятности ошибок обоих родов зависят от величины порога: чем ниже порог, тем больше вероятность правильно принять посылку тока, то есть тем меньше уровень значимости, но зато возрастает вероятность ошибки второго рода — ошибочно принять паузу за посылку тока. Таким образом, при обсуждаемом методе проверки выбор порога — это выбор вероятности ошибок обоих родов.

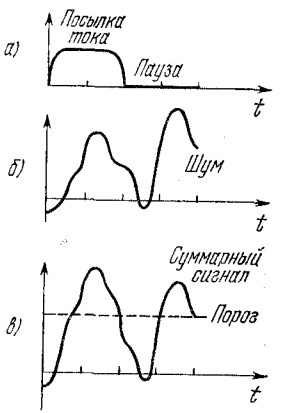


Рис. 3.

В радиолокации или при контроле качества продукции ситуация аналогична. Как же связаны вероятности ошибок первого и второго рода? Допустим, мы хотим обеспечить очень маленькую вероятность ложной тревоги в системе обнаружения ПВО, ну, например, не более одной ложной тревоги на десять миллионов импульсов, то есть выбрать уровень значимости $\alpha=10^{-7}$. Но такие вещи даром не даются — придется расплачиваться. Доведем рассуждение до логического конца: обратим в нуль уровень значимости. При этом придется все сигналы независимо от реального положения дела трактовать как шум. Теперь оператор никогда не объявит ложной тревоги, но он же вообще не объявит тревоги, даже когда армада самолетов будет у него над головой! Такой план проверки, мягко выражаясь, нельзя считать удовлетворительным.

Результат нашего парадоксального рассуждения указывает на необходимость тщательного сопоставления последствий ошибок первого и второго рода и выбора какого-то компромисса.

Казалось бы, в математической статистике должны быть разработаны методы для назначения или вычисления допустимых вероятностей ошибок при проверке гипотез. Однако *я должен огорчить вас, читатель; эти вопросы* лежат не только вне математической статистики, но пока вообще вне математической науки. Допустимая величина уровня значимости — это же мера риска лица, принимающего решение, если сейчас мы оцениваем всю ситуацию с его позиций.

Подробный разговор о риске у нас еще впереди, но сейчас хочу обратить ваше внимание на крайнюю субъективность допустимого уровня риска. Если вы когда-нибудь играли в карты или другую игру на деньги, то знаете, как поведение игрока зависит от величины возможного выигрыша и проигрыша и характера игрока — его азартности. А в задачах, связанных с техникой или природой, оценки допустимого риска оказываются зависящими еще от многих других причин, например таких, как престиж или квалификация.

Положение все же не столь безнадежно, как может сейчас вам показаться. Мы пока обсуждали лишь самый простой метод обработки поступающих сигналов. Когда при разговоре по телефону вас плохо слышат, вы повторяете одно и то же несколько раз. Так же и в радиолокации: можно посылать не один импульс для обнаружения самолета, а несколько подряд, как говорят, пакет импульсов. Теперь, имея пакет, можно при проверке гипотезы о наличии сигнала с шумом (H_1) или только шума (H_0) воспользоваться различными методами обработки информации.

Будем записывать 0, когда принято решение *шум*, и 1, когда принято решение *сигнал с шумом*. Пусть в пакете сто импульсов, и относительно каждого сделано заключение, есть ли это 0 или 1, то есть пакет импульсов после обработки — это сто чисел, каждое из которых есть 0 или 1. Можно, например, принять такое правило: если среди ста символов количество единиц больше пяти (а нулей соответственно меньше 95), то принимается решение о наличии сигнала, в противном случае принимается решение о наличии только шума. Если вместо пяти взять любое другое число, получится иное правило проверки. Словом, ясно, что вероятности ошибок обоих родов будут зависеть от количества импульсов в пакете, допустимого количества единиц при принятии решения об отсутствии сигнала и, конечно, от вероятности правильного обнаружения каждого импульса. Таким образом, в распоряжении исследователя довольно много возможностей, но нужно хорошо подумать о том, как их использовать.

Когда приятель приглашает вас принять участие в лодочном походе, вы можете его обрадовать, сказав «ДА», или сказать «НЕТ» и повергнуть его в уныние... Но у нас есть еще один ход — «МОЖЕТ БЫТЬ».

Велико разнообразие мотивов, которые приводят к выбору одного из этих ответов, не менее велико разнообразие

последствий. Но обратите внимание на третий из возможных ответов, резко отличающийся от первых двух своей неопределенностью. Говоря «МОЖЕТ БЫТЬ», вы оттягиваете окончательное решение. И делаете это неспроста: вы недостаточно знаете других попутчиков или неясно представляете себе маршрут, вам неизвестна материальная сторона или не согласовано время отпуска. Итак, вы говорите «МОЖЕТ БЫТЬ» из-за недостатка информации, нужной для принятия окончательного решения.

При статистической проверке гипотез мы находимся в той же обстановке: когда имеющаяся информация оказывается недостаточной для решительного «ДА» или «НЕТ», то есть для принятия или отвержения гипотезы, тогда мы можем отложить решение — сказать «МОЖЕТ БЫТЬ» — и принять необходимые меры для получения недостающей информации.

Только что мы обсудили метод проверки гипотез в радиолокации, при котором количество импульсов в пакете фиксировалось, и решение о принятии гипотезы об отсутствии сигнала или ее отвержении принималось после того, как становились известными результаты обработки всех импульсов в пакете. Легко усмотреть слабости такого метода. Если принимается решение о наличии сигнала, когда в пакете оказалось больше пяти единиц, и среди первых десяти импульсов оказалось семь единиц, то независимо от остальных девяноста следует принимать решение о наличии сигнала. Таким образом, время и мощности, затраченные на обработку девяноста импульсов, тратятся зря. Если же среди первых семидесяти импульсов нет ни одной единицы, то можно было бы поверить в отсутствие сигнала и не проверять остальные тридцать.

Вот здесь и возникает идея не всегда после обследования пакета принимать решение об отсутствии или наличии сигнала, то есть о принятии или отвержении нуль-гипотезы. Если число нулей в пакете мало, то разумно гипотезу о наличии сигнала принять, а если нулей много, столь же обоснованно принять нуль-гипотезу. Но когда число нулей в пакете не очень мало и не очень велико, то естественно считать имеющуюся информацию недостаточной для принятия решения, то есть вместо «ДА» или «НЕТ» принять решение «МОЖЕТ БЫТЬ» и провести дополнительные наблюдения.

Простейшая процедура такого типа — план из двух последовательных пакетов.

Разберем пример. При проверке гипотезы об отсутствии сигнала, допустим, берется пакет из сорока импульсов. Если среди них нет ни одной единицы или есть единица, а остальные числа — нули, то принимается гипотеза об отсутствии сигнала. Если среди сорока импульсов окажется шесть единиц или больше, то принимается гипотеза о наличии сигнала. Если же окажется две, три, четыре или пять единиц, то посылается повторный пакет, скажем, из тридцати импульсов. Теперь нужно назначить новое допустимое число единиц, но уже по суммарному числу импульсов, то есть по семидесяти. В нашем примере его можно принять равным семи. Таким образом, если в объединенном пакете среди семидесяти импульсов окажется менее семи единиц, принимается гипотеза об отсутствии сигнала, а если окажется семь единиц или больше, принимается альтернативная гипотеза о наличии сигнала. Та же самая ситуация, конечно, складывается при выборочном контроле качества продукции.

Подсчеты подтверждают, что такие планы с повторной выборкой эффективнее одновыборочных планов: в среднем в выборке участвует меньшее количество импульсов (или изделий) при тех же результатах, либо получаются лучшие результаты — понижаются вероятности ошибок обоих родов при том же среднем количестве импульсов.

Возможно у вас сейчас осталось какое-то ощущение незаконченности: почему надо ограничиться двумя последовательными выборками, а не сделать их три, четыре, или, вообще, столько, сколько нужно?

Так оно и есть в действительности. Но одно дело — нечеткие соображения и совсем другое — создание далеко идущей теории.

Даже для задач выборочного контроля уже ясно, что можно предложить множество различных планов. Если иметь в виду широкий круг проблем статистической проверки гипотез, то, думаю, вам ясна возможность построения самых различных планов, и здесь для фантазии создателя плана открыт широкий простор: можно по-разному ставить задачи проверки гипотез, и каждой будет соответствовать даже не один, а множество планов.

Компромисс

Итак, выбор плана или процедуры проверки гипотез во власти исследователя, и теперь ему нужно выбрать наилучший план. Мы уже знаем — прежде всего нужно указать критерий качества процедуры проверки.

Зададим уровень значимости (вы помните — это вероятность ошибочно отвергнуть нуль-гипотезу, когда она верна), уровень, который устраивает исследователя в данной ситуации. Так как планов теперь множество, то, вообще говоря, различными будут соответствующие им ошибки второго рода. А исследователю, конечно, хочется по возможности уменьшить вероятности ошибок обоих родов. Поэтому при заданном уровне значимости за критерий качества плана можно взять величину вероятности ошибки второго рода.

В этой ситуации имеется возможность поставить задачу оптимизации: при заданном уровне значимости выбрать процедуру принятия решения, при которой вероятность ошибки второго рода будет наименьшей.

Сформулированный принцип выбора оптимальной процедуры составляет центральную идею одного из методов статистической проверки гипотез, развитого выдающимися зарубежными статистиками Ю. Нейманом и Е. Пирсоном в середине тридцатых годов. Можно доказать — и это утверждение играет значительную роль в современной мате-

матической статистике, — что при заданном уровне значимости такая процедура существует. Она основана на отношении правдоподобия и может быть реализована.

Критерий Неймана — Пирсона широко применяется в математической статистике и, в отличие от достаточно произвольных критериев оценки сочинения на литературную тему, которые я привел в разделе «Немного о критериях», имеет, как вы, надеюсь, заметили, достаточно обоснованный характер.

Однако на самом деле в задаче статистической проверки гипотез, как и в любых других областях человеческой деятельности, критерий не единствен, могут быть предложены и другие, тоже вполне обоснованные. В то же время в некоторых ситуациях критерий Неймана — Пирсона весьма уязвим... Вот об этом и пойдет сейчас речь.

Мне кажется полезным еще раз обсудить саму меру доверия, хотя о ней уже неоднократно говорилось. Как мы уже установили, при выбранной процедуре проверки гипотез за понижение уровня значимости приходится расплачиваться увеличением вероятности ошибки второго рода, то есть — при выборочном контроле — потерями за счет браковки годной продукции.

И здесь, как я уже говорил, возникает весьма острая проблема компромисса между интересами поставщика и потребителя. Этот компромисс требует тщательного изучения последствий, к которым приводят ошибки первого и второго рода, учета наносимого ущерба.

Для участников компромисс всегда труден — каждый, даже понимая необходимость уступок, подсознательно, а то и совершенно сознательно борется за уменьшение грозящих ему потерь — не случайно же они названы потерями.

Ежедневно всем приходится как-то согласовывать свои поступки с начальством, подчиненными и коллегами, с членами семьи и соседями, словом, жить в обществе. Согласование в лучшем случае означает взаимное согласие, но интересы участников могут не только не совпадать, но и быть противоположными.

Когда вечером в пятницу супруги собираются в гости к друзьям в соседний дом, и Она уже соответствующим образом подготовилась, возникает полемика. Ему хотелось бы пойти в тренировочном костюме и тапочках: смотреть футбол, играть в преферанс или пить чай с вареньем можно и так... Но Она настаивает на черном костюме, крахмальной сорочке и новых полуботинках — нужно же показать этой неряхе — соседке, как следует содержать мужа. Но новые туфли не разношены и чуть-чуть жмут (совершенно достаточно для создания отвратительного настроения), а мысль о твердом воротничке с галстуком вызывает раздражение. После препирательств и очевидных взаимно ободряющих эпитетов и метафор Он надевает костюм, шерстяную рубашку без галстука и вполне еще приличные разношенные туфли. Таким образом, даже в хорошо налаженной семейной жизни интересы супругов подчас совершенно не совпадают.

Когда же умиротворенные супруги, наконец, появляются, то проницательная хозяйка дома сразу замечает слишком уж помпезный вид гостьи и тут же строит гипотезу об ее интимном интересе к неожиданно посетившему их сослуживцу.

Вот здесь уже интересы подруг хотя и не совпадают, но, очевидно, не противоположны: одна просто любопытна или хочет иметь в руках козырь, а другая, если гипотеза верна, вовсе не хочет открывать карты.

Итак, нуль-гипотеза хозяйки дома *интимный интерес есть*. Ошибка первого рода — отвержение этой гипотезы, когда на самом деле у гостьи есть интимный интерес к сослуживцу, — хозяйке ничем не грозит, хотя и принесет женщине, любящей сенсации, некоторое разочарование. Но ошибка второго рода — принятие гипотезы, когда на самом деле никакого «интереса» нет, грозит подруге сплетнями, от которых подчас так трудно избавиться!

В то же время, если гипотеза об интимном интересе на самом деле верна, то гостья знает, на что идет, и ей легче предусмотреть последствия и принять меры, если хозяйка догадается о наличии интимного интереса со всеми вытекающими отсюда последствиями.

Вот здесь и возникает компромиссное решение — гостье нужно выбрать стратегию поведения, дающую, с одной стороны, возможность реализовать удачное (или подготовленное!) стечение обстоятельств и, с другой, не открывать карты.

Неприятности, которые грозят дома, как правило, носят моральный, а не материальный характер, и поэтому ни участница событий, ни посторонний наблюдатель, которому безразлично их течение, не смогут назначить допустимую количественную меру для ошибок, или, как говорят, назначить цены ошибок.

События текут своим чередом, и после милого чаепития мужчины садятся играть в преферанс. Вот здесь уже интересы взявшего прикуп точно противоположны интересам объединившихся — открывших карты противников, ибо выигрыш ведущего игру — это проигрыш его противников.

Но вернемся к интересам поставщика и потребителя, где не только очевидна необходимость компромисса при оценке возможных потерь из-за ошибок обоих родов, но и имеется возможность ввести численную меру.

Однако, несмотря на очевидность компромисса, в реальных условиях ввести численную меру совсем не легко. Дело в том, что нужно каким-то разумным способом задать допустимую в данной ситуации вероятность ошибки. А вот как указать этот способ, на основании каких соображений можно предложить принцип для построения правила, по которому будет назначаться допустимая вероятность ошибки,— все это вопросы, на которые сегодня нет ответа. И мне они представляются крайне сложными.

Попробую пояснить возникающие здесь трудности.

Обратимся вновь к проблеме обнаружения в радиолокации. Напомню: служба обнаружения достаточно сложная и дорогостоящая, но неизмеримо дороже может оказаться ее ненадежная работа. Скажем, если речь идет о корабельном радиолокаторе, то ночью или в тумане его задача — предотвратить столкновение с другим кораблем, айсбергом или подводной скалой, столкновением, результатом которого могут быть не только огромные материальные потери, но и гибель сотен людей.

Вот и попробуйте назначить величины вероятностей ошибок первого и второго рода, которые могут вас устроить! Скажем, вы назначаете вероятности ошибок по одной тысячной. Что это означает? Если сигналы поступают, например, каждую сотую долю секунды (то есть сто сигналов в секунду), то вероятность ошибки первого рода в одну тысячную означает, что в среднем за тысячу сигналов будет один пропуск, то есть в среднем один пропуск за десять секунд, то бишь шесть пропусков в минуту. Это еще в среднем! А ведь можно пропустить сразу и несколько сигналов... Служба обнаружения с такой низкой надежностью просто никуда не годится. Если же назначить вероятности ошибок по одной миллионной при тех же ста сигналах в секунду, то в среднем будут приниматься ошибочные решения один раз за десять тысяч секунд, то есть

примерно за три часа... А нам хотелось бы вообще не иметь подобных ошибок!

Но, простите, безошибочных систем не бывает... Как же нам быть?

Ну, тогда желательно иметь систему, хотя бы столь надежную, чтобы в течение нашей быстротекущей жизни не было таких трагических ошибок. И в системах обнаружения приходится усложнять аппаратуру, платить за нее колоссальные суммы, содержать огромный штат для обеспечения высокой надежности этих систем.

Пример с системой радиолокационного обнаружения для многих сложен, ибо далеко не все дети, женщины, а подчас и мужчины представляют себе эту систему. Поэтому приведу другие, более близкие каждому примеры.

Пешеход, переходя улицу, рискует, даже когда он точно соблюдает правила: их может нарушить шофер или бегущий подросток, нечаянно толкнувший вас на мостовую.

Газета «Вечерняя Одесса» ежедневно сообщала, сколько за сутки произошло дорожно-транспортных происшествий, аварий и наездов на веселых и беспечных одесских пешеходов: в сутки в среднем два человека получали увечья или гибли. Если принять, как это и следует сделать, частоту за оценку вероятности — здесь, к сожалению, данных предостаточно, то оказывается в Одессе — городе с миллионным населением — вероятность ежедневно попасть в подобную беду для каждого есть величина порядка 0,000002 — далеко не такое уж малое число, если это касается вашей жизни или здоровья ваших детей.

Любая мать с возмущением отметет самую постановку вопроса о назначении допустимой для нее вероятности ошибки первого рода, то есть несчастного случая, когда она переходит улицу и тянет за собой размазывающего слезы и упирающегося карапуза или посылает подростка в магазин через дорогу. Психология матери допускает единственную вероятность несчастного случая — нуль! И попробуйте ей объяснить, что этого можно добиться, лишь вообще не выходя на улицу, не оставляя детей одних дома, не живя в доме, ибо может быть пожар у соседей или обрушиться дом, словом, имеется сколько угодно несчастий, но, к счастью, они происходят редко, с малой вероятностью.

Вот и непонятно, какими принципами следует пользоваться, назначая допустимую вероятность ошибки первого

рода для несчастий, которые непосредственно с вами могут произойти. На каких принципах должны основываться решения горсоветов, когда они назначают предельную скорость движения автотранспорта по городу, повышая ее, скажем, с пятидесяти до шестидесяти километров в час? Пешеходы-то полагают, что чем быстрее движется транспорт, тем больше вероятность аварии. На самом деле это не так: повышение предельных скоростей подразумевает и изменение организации движения, соответствующее воспитание пешеходов и водителей, ужесточение мер наказания за нарушение правил дорожного движения этими группами населения, каждая из которых считает, что именно другая должна соблюдать и не нарушать...

В действительности никто и не назначает эту вероятность, пешеходы привычно сетуют на милицию, не обеспечивающую безопасность на улице, и играют в поддавки с несущимися автомобилями, а городские власти вынуждены увеличивать скорость движения, исходя из совершенно других соображений, в существе которых содержится забота и о пешеходе, — он подчас тоже едет на автомобиле, а кроме того, ему ежедневно нужна пища и телесная, и духовная, вот ее-то и приходится доставлять.

Правила дорожного движения, обязывающие пешеходов и водителей транспорта быть внимательными и осторожными, призваны понизить вероятность ошибки, которая может привести вас на больничную койку или даже в «лучший» мир, в который торопиться не следует. Здесь это есть ошибка первого рода при обычно принимаемой нуль-гипотезе: *при переходе улицы лично со мной ничего трагического не произойдет*. В то же время здесь как раз ошибка второго рода — ложная тревога вполне допустима: она же означает, что вы лишь подождали зря, не перебежали перед носом троллейбуса, остерегаясь автомобиля, который мог выскочить из-за него, но не выскочил — водитель соблюдал осторожность.

На самом же деле мы спокойно переходим улицу, хотя и знаем о возможности аварии или несчастного случая, исходя из принципа практической уверенности, четкую формулировку которого я впервые встретил в учебнике по теории вероятностей академика С. Н. Бернштейна.

Принцип практической уверенности: если вероятность события мала, то следует считать, что в однократном эксперименте это событие не произойдет.

Впрочем, еще П. Л. Чебышев в 1845 году в магистерской диссертации «Опыт элементарного анализа теории вероятностей» писал: «Приближенно мы считаем несомненным, что события будут или не будут иметь место, если вероятности их мало отличаются от 1 или 0».

Вот этот принцип и дает человечеству возможность жить, не находясь в постоянном страхе по поводу всех трагических событий, которые хотя и могут произойти, но, к счастью, с малой вероятностью.

Однако что означает — с *малой вероятностью*? Все тот же вопрос: одна сотая — уже малая? Или нужна одна миллионная, или еще меньше?

При нормальном режиме работы электронной лампы, стоящей в вашем телевизоре, вероятность перелета некоторого вполне определенного электрона с катода на анод в течение одной секунды есть величина порядка одной миллиардной. Казалось бы, никакой ток через лампу не должен идти. Но электронов много, в течение секунды перелетает с катода на анод примерно 10^{16} электронов, и поэтому величина вероятности перелета 10^{-9} не представляется такой уж маленькой.

Нечто аналогичное происходит и в городе.

Хотя вероятность именно вам оказаться участником дорожно-транспортного происшествия и мала, но в городе много людей, и вероятность хотя бы одного, а то и нескольких происшествий уже не мала, а подчас весьма близка к единице, и поэтому, к сожалению, так часто несутся по городу кареты скорой помощи...

Понятие малости при оценке вероятности индивидуально. Уверенность в своих силах, характерная для подростков, повышает эту границу, хотя и здесь есть свои пессимисты и оптимисты. Но, кроме того, оценка вероятности зависит от ситуации.

Я приведу пример подобной ситуации.

По наблюдениям медиков в США процент больных раком желудка меньше, чем в Европе. Это наблюдение связали со способом употребления спиртного: американцы пьют крепкие спиртные напитки разведенными (знаменитые виски с содовой или джин с тоником), в то время как европейцы пьют неразведенные.

Легко себе представить две группы исследователей, одна из которых предпочитает крепкие напитки и хочет доказать их безвредность или вредность, хотя бы превосходящую результаты от потребления разведенных напит-

ков, а другая группа предпочитает разведенные напитки и ставит своей целью доказать их заметно меньшую вредность. Обе группы вроде бы защищают интересы пьющих, но вторая группа при этом отстаивает интересы фабрикантов безалкогольных напитков, таких, как содовая вода или тоник.

Назовем группы исследователей соответственно крепкой и разведенной. Их задача состоит в проверке гипотезы, принимаемой за нулевую,— *среди заболевших раком желудка доля пьющих крепкие напитки равна доле пьющих разведенные.* Здесь альтернативная гипотеза: *среди заболевших раком желудка доля пьющих крепкие напитки больше доли пьющих разведенные.*

Казалось бы постановка задачи в такой форме должна приводить к отчетливому ответу: либо верна нуль-гипотеза, либо неверна. Однако это не всегда так. Группа крепких хочет доказывать верность нуль-гипотезы, и поэтому ей выгодно выбрать такую процедуру проверки, при которой нуль-гипотеза отклонялась бы весьма редко, то есть ей выгодно уровень значимости выбрать поменьше. Ошибка второго рода их мало интересует: она означает, что принимается устраивающая группу крепких гипотеза о равной вредности, когда на самом деле крепкие напитки вреднее разведенных.

В то же время группа разведенных заинтересована в противоположной трактовке тех же наблюдений, и поэтому ей выгодно выбрать поменьше вероятность ошибки второго рода, то есть, чтобы альтернативная гипотеза о сравнительно большем вреде крепких напитков отвергалась как можно реже. Но при этом группа разведенных заинтересована в том, чтобы вероятность отвергнуть нуль-гипотезу, если она ложна, была очень большой.

Вследствие столь разного подхода или понимания важности ошибок первого и второго рода эти группы исследователей могут выбрать существенно различные процедуры принятия решения, а потому и разные ответы на вопрос.

Мой близкий друг — известный врач рассказал мне об этой проблеме, которую он почерпнул в специальной литературе.

Оказалась статистически достоверной альтернативная гипотеза, так что группа разведенных, а вслед за этим и американский образ употребления крепких напитков восторжествовали. Однако группа крепких не успокоилась, и им удалось статистически достоверно доказать, что при

заболевании раком толстой кишки наблюдается обратная картина: он чаще возникает у любителей разведенных напитков. Казалось, это должно было бы примирить спорящих. Но сегодня хирургическое вмешательство и лечение дают лучшие результаты при раке толстой кишки, и пока вновь преимущество на стороне группы разведенных.

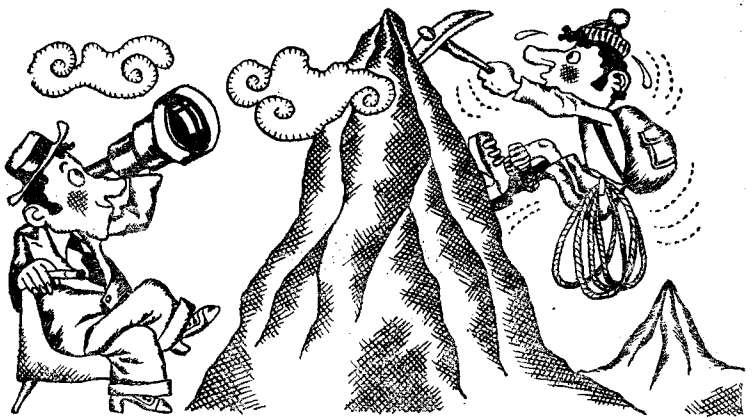
Все, конечно, знают, сколь большое распространение получил рак, и не кажется ли вам, читатель, странным, что подвергается пристальному изучению острая проблема: *кто страшнее: змий крепкий или змий разведенный?* вместо проверки нуль-гипотезы *всякий змий вреден* и, при положительном ответе, а мне он представляется почти очевидным, разработки эффективных мер по борьбе со змием любой крепости?

Вернемся к теории Неймана — Пирсона. Проведенное в этом разделе обсуждение, как мне кажется, убедительно показывает не столь уж значительную, как это представляется с первого взгляда, обоснованность их подхода.

В самом деле в теории Неймана — Пирсона критикуется прежде всего достаточно произвольный выбор одной из двух альтернативных гипотез, который царил в статистике до их теории. Однако разработанная ими теория, о которой я кратко рассказал, переносит этот произвол на выбор допустимого уровня значимости. Конечно, после того как уровень значимости выбран, проверка гипотез производится со всей научной строгостью. Но значительный произвол все равно остается, он просто перенесен глубже, и теперь труднее обнаружить, сколь велики последствия такого произвола.

Не подумайте, будто эти замечания — повод отказаться начисто от принципов, положенных в основу теории Неймана — Пирсона: есть задачи, где допустимый уровень значимости указать можно и вполне разумным образом. Более того, в некоторых ситуациях удается получить даже аналитическую зависимость между вероятностями ошибок первого и второго рода, и исследователь имеет возможность четко увидеть, чем он должен расплачиваться за уменьшение уровня значимости.

Критические замечания в адрес теории Неймана — Пирсона скорее говорят о том, что и после их важных исследований в теории проверки гипотез проблема не была закрыта. И сейчас, после интенсивных исследований последних десятилетий, теория статистической проверки гипотез все еще далека от завершения.



Динамика вместо статики

Вернемся еще раз к проблеме статистической проверки гипотез в задаче радиолокационного обнаружения. Ломать — не строить, и не удивительно, что мне легко удалось развенчать процедуру проверки, в которой используется один пакет импульсов фиксированного объема. Хотя план из двух пакетов лучше, но, как вы заметили, можно пойти и дальше, рассмотреть планы из трех или четырех пакетов (или выборок), но все они имеют один фундаментальный дефект — статичность.

Такой подход противоречит нашему жизненному опыту: прежде чем принять решение при покупке нового костюма, поисках нового места работы или при вступлении в жилищный кооператив, вы же присматриваетесь, прицениваетесь, выясняете многочисленные «за» и «против» и, как правило, заранее не фиксируете время или количество наблюдений, достаточные для принятия решения.

Процесс воспитания ребенка выглядит аналогично: вы имеете возможность либо его поощрить за хорошее поведение — купить мороженое, либо наказать — запретить посещение кино. Ребенок все время что-то совершает, но вы, наблюдая за его поступками, откладываете принятие решения либо до того момента, когда переполнится чаша терпения и придется прибегнуть к наказанию, либо переполнится чаша умиления, и вы решитесь на поощрение.

Почему же, когда дело касается контроля качества продукции, радиолокационного обнаружения или другой из обсуждавшихся задач, нужно действовать иначе, фиксируя заранее объем выборки? Ведь если при проверке заранее фиксируется объем выборки, то последовательность действий вовсе не зависит от постепенно накапливающихся данных.

Таким образом, видно, что одновыборочный план построен не рационально. При процедуре контроля из двух последовательных выборок действия хотя и зависят от результатов первой выборки — не всегда надо делать повторную выборку, — но все же зависят весьма слабо.

Поэтому на смену статичности приходит идея последовательного анализа. Его главное отличие от планов с фиксированным объемом выборок именно в том, что само количество наблюдений не фиксируется заранее — оно зависит на каждой стадии наблюдений от предыдущих результатов и, таким образом, является случайной величиной.

Казалось бы, это простая идея — не фиксировать заранее время или число наблюдений, а принимать решение после получения нужной информации — лежит на поверхности, и непонятно, почему сразу статистики к ней не обратились. Впрочем, если вдуматься поглубже, то далеко не все так уж очевидно.

Прежде всего без скрупулезного подсчета вовсе не ясны преимущества последовательного плана по сравнению с одновыборочным или с более эффективным двухвыборочным, которые уже освоены. Однако главное, конечно, не в этом: последовательный, динамический анализ в задачах проверки гипотез — это новый подход.

Двухвыборочный план контроля — прогресс по отношению к одновыборочному — был разработан в конце двадцатых годов. Затем лишь в сороковых годах начали изучать многоступенчатые выборки и некоторые последовательные процедуры. Но основное продвижение здесь принадлежит американскому математику Абрахаму Вальду, развившему общие идеи последовательного анализа в годы второй мировой войны. Они стимулировались военными организациями, и разработанные методы сразу же были применены к конструкторским работам в области военного и военно-морского оборудования.

Кардинальные открытия всегда просты, но они требуют небанального, принципиально нового взгляда на изве-

стные факты, преодоления установившихся точек зрения, канонов, привычек и поэтому требуют недюжинного таланта, а то и мужества — на пути новой идеи обычно стоит высокий забор из старых и привычных представлений.

Скажем, идея фламастера — ручки, в которой чернила подаются из баллона по тоненькой трубочке со вставленным внутрь гигроскопическим стержнем, например из фетра, очевидно, проще, чем идея металлического аналога гусиного пера — знаменитого пера № 86 или его не известного мне предшественника. Однако примерно столетие дети и взрослые макали это проклятое перо в чернильницы, мазали руки, да и не только руки, мучительно старались выучиться никому не нужному нажиму при каллиграфическом письме.

Прежде чем сообщить вам о выгодах, которые несет последовательный анализ по сравнению с традиционными методами, я подробнее расскажу о самом методе.

Спортивные танцы на льду или классический балет требуют от балерины многого и, в частности, сохранения оптимального веса. Увеличение веса даже на пару килограммов приведет к разладке не только собственных рефлексов, но и нарушит привычное согласование с партнером, так что он, если и не уронит партнершу, но не удержит ее нужное время или бросит не на точно рассчитанное расстояние, и слаженность, а с ней и красота, будут нарушены. Обычно танцовщица держит минимально возможный вес, и поэтому проблему похудения я обсуждать не буду.

Для простоты будем рассматривать ситуацию, когда есть всего две гипотезы: H_0 — *вес танцовщицы равен* 50 *килограммам* и H_1 — *вес равен* 48 *килограммам.*

Представьте себе, что танцоры отправились на длительные гастроли, где, конечно, привычный режим нарушен, а танцовщица должна по-прежнему держать свои 48 килограммов. И возникает задача слежения за весом: нужно во время заметить его увеличение и принять меры — изменить режим питания и тренировок. Но вес человека — величина отнюдь не стабильная, он меняется на сотни граммов даже в течение дня, да и измеряется с ошибкой. Поэтому нужно производить не одно измерение, а повторять их через какие-то интервалы времени и следить за средним. Но сколько же раз измерять и как часто? Два-три может оказаться мало, а сто измерений не пре-

ведешь... Вот здесь и приходит на помощь последовательный анализ.

Итак, танцовщица борется за свои 48 килограммов. Ошибки измерения — случайные величины. Но есть и другие причины, вызывающие хотя и не очень значительные, но заметные случайные колебания веса, — не всегда лучше переесть, чем недоспать...

Вся ситуация призывает нас обратиться к отношению правдоподобия, но только с некоторой модификацией. Задавшись вероятностями ошибок первого и второго рода (в этой задаче их можно принять одинаковыми), логично задать два порога и установить такое правило: если отношение правдоподобия превосходит больший из порогов, принимается нуль-гипотеза — *вес танцовщицы равен 50 килограммам*, если оно оказывается меньше другого, принимается альтернативная гипотеза — *вес оказался равным 48 килограммам,* и, наконец, если величина отношения правдоподобия заключена между порогами, то нет еще основания для принятия любой из гипотез и следует продолжать наблюдения.

Я все время говорил о проверке нуль-гипотезы или альтернативной и о принятии одной из них. Математики часто предпочитают другую интерпретацию, и вместо принятия альтернативной гипотезы говорят об отвержении нуль-гипотезы.

При такой трактовке очень удобно говорить «ДА» при принятии нуль-гипотезы и «НЕТ» при ее отвержении. Если же нельзя сказать ни ДА, ни НЕТ — не хватает информации, то следует сказать МОЖЕТ БЫТЬ и продолжить наблюдения.

Всю обстановку очень наглядно можно представить графически, но для ее пояснения придется написать кое-какие формулы.

Последовательные значения измеренного веса танцовщицы обозначим $x_1, x_2, \ldots, x_n$. Отношение правдоподобия — функция от этих переменных.

Распределение вероятностей величины веса человека можно обычно считать нормальным. Однако если верна нуль-гипотеза, то математическое ожидание веса будет 50 килограммов, а при альтернативной гипотезе — 48 килограммов. В этих условиях проверка того, находится ли отношение правдоподобия в границах, заданных порогами, или оказывается за одной из этих границ, сводится, путем несложных выкладок, к проверке совсем простых

неравенств

$$a + n \cdot 49 \leqslant \sum_{i=1}^{n} x_i \leqslant n \cdot 49 + b, \qquad (*)$$

где n — количество измерений — текущая переменная, пропорциональная времени наблюдения. Число 49 появилось здесь как полусумма границ 50 и 48 килограммов, а числа a и b зависят от назначенных заранее величин вероятностей ошибок первого и второго рода.

Если вместо 50 и 48 килограммов записать неравенства в общем виде, обозначив полусумму нижней и верхней границ допустимого изменения измеряемой переменной через μ, то вместо неравенства $(*)$ следует написать:

$$a + n\mu \leqslant \sum_{i=1}^{n} x_i \leqslant n\mu + b. \qquad (**)$$

Эти неравенства можно теперь изобразить графически (рис. 4): в координатах (n, x) правая и левая части — это прямые, а $\sum x_i$ представляет собой ломаную. Когда ломаная пересекает верхнюю прямую, то она попадает в область, отвечающую нуль-гипотезе, и принимается решение «ДА», когда ломаная пересекает нижнюю прямую, она попадает в область отвержения нуль-гипотезы, и принимается решение «НЕТ». А до тех пор, пока с возрастанием числа наблюдений ломаная движется внутри полосы, принимается на каждом шаге решение «МОЖЕТ БЫТЬ» и производится следующее измерение.

Проблема сохранения веса танцовщицы, возможно, не имеет столь уж большого общественного значения. Но мы можем вновь обратиться к контролю качества серийной продукции. Например, когда работает станок-автомат и измеряемый параметр — это размер детали, который должен не выходить за пределы допуска, то вся ситуация абсолютно совпадает с уже разобранной. Нужно лишь заменить слова «танцовщица» на «деталь», вместо веса будет размер, нуль-гипотезе отвечает выход за пределы допуска. Выход за пределы допуска — это брак, и следовательно, нужно остановить станок для подналадки.

Когда вероятности ошибок первого и второго рода — величины, не слишком сильно различающиеся, то, как показал еще А. Вальд, последовательная процедура по сравнению с традиционной, когда количество испытаний заранее фиксируется, даст сокращение числа испытаний в среднем примерно вдвое. При контроле качества продукции уменьшение числа контролируемых изделий вдвое — это значительный выигрыш, особенно если продукция дорогостоящая или если сами испытания занимают много времени и средств. Поэтому последовательный анализ широко используется при контроле качества продукции.

Но есть еще один круг задач, где применение последовательного анализа оправдано дает значительно больший выигрыш.

При радиолокационном обнаружении для повышения надежности посылают сигналы многократно, периодически их повторяя, и суммируют принятые сигналы. При этом нерегулярные шумы, принимая по воле случая то положительные, то отрицательные значения, при суммировании в среднем будут уменьшаться относительно суммарного сигнала.

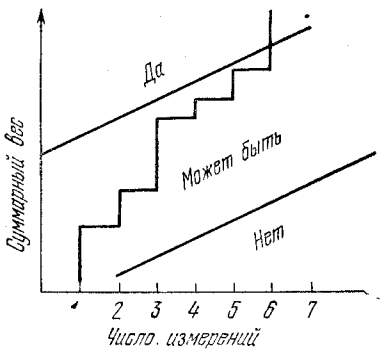
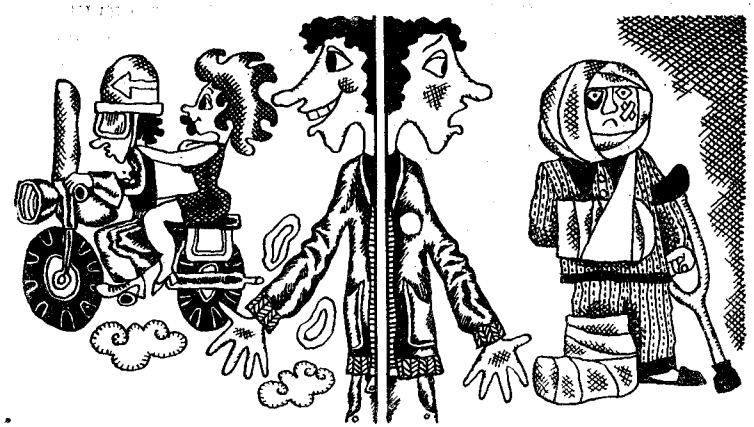


Рис. 4.

При радиолокационном обнаружении, как мы это уже обсуждали, ошибки первого и второго рода, то есть пропуск сигнала (а вместе с ним и обнаруживаемого объекта) и ложная тревога далеко не равноценны. И если считать их величины существенно различными (все равно, какая из них много меньше другой), то последовательная процедура при обработке принимаемых сигналов дает весьма значительный выигрыш в количестве наблюдений, а вместе с тем как во времени обнаружения, так и в потребной мощности излучений по сравнению с традиционной процедурой, при которой фиксируется количество наблюдений. При необходимости принимать срочные меры сокращение времени обнаружения дает весьма существенные преимущества, а уменьшение потребной мощности эквивалентно увеличению дальности, на которой станция уже может обнаружить тот самый нежелательный объект. Конечно, и здесь есть свои подводные камни: уменьшение количества наблюдений происходит лишь в среднем, но дисперсия здесь довольно велика. Поэтому применение последовательного анализа следует осуществлять после тщательного расчета. Но я не буду здесь останавливаться на этих тонкостях.



Риск

Двое современных парней Игорь и Михаил отправляются в отпуск: имеются путевки в пансионат на Кавказ. У одного есть автомобиль, у другого — мотоцикл. Кроме того, можно отправиться на поезде или самолете. Выбор велик и разнообразен. На самолете 2 часа полета, и они уже вблизи моря... Но жена Игоря возражает против самолета: ей приснилась авария, и она умоляет их ехать поездом. Впрочем, оба хотели бы иметь на Кавказе автомобиль или мотоцикл: перспектива просидеть на пляже 24 дня с одними и теми же курортниками их не устраивает, хочется поездить, посмотреть. Но если отправляться в путешествие, автомобиль требует серьезного ремонта, а времени остается очень мало. На мотоцикле же проехать полторы тысячи километров для Игоря не такое уж большое удовольствие: говорят, что на мотоцикле следует ездить до 25 лет, пока хорошо срастаются кости... А Игорю уже перевалило за 35. Но Михаил как раз в самом «мотоциклетном» возрасте — 21 год, и ему не так уж хочется на Кавказ, как гложет детское желание «жжж... быстрее всех».

Если они едут на автомобиле или мотоцикле, то отпуск наступает, так сказать, немедленно по выезде из дома, и потерь времени на переезд практически нет. Но, конечно, возможны другие неприятности: прежде всего, на мотоцикле больше шансов попасть в аварию, чем на автомобиле, на поезде или самолете. Впрочем, авария на автомобиле — большей частью помятый кузов, в то время как на поезде — это, в лучшем случае, переломы и ушибы, ну а на самолете — сами понимаете. Но дело не только в опасности езды на мотоцикле: сидеть, особенно сзади без спинки, расставив неудобно ноги, в неизменной позе по многу часов в день — удовольствие не слишком большое... Наконец, надо понять и внешнюю сторону: как шикарно в небрежной позе сидеть на мотоцикле перед входом в пансионат, надев потрепанную кожанку и положив рядом красно-белый шлем, когда разморенные курортники в домашних тапочках тащатся с пляжа с детьми и авоськами... Каждому видно, как у девушек захватывает дух. Нет, без мотоцикла нельзя. И между друзьями идет многодневное обсуждение наилучшего варианта. Теперь обсудим всю ситуацию с научных позиций.

Едва ли нужно объяснять причины желания хорошо провести отпуск: мы же не зря прилагаем усилия и тратимся сверх меры, лишь бы осуществить эту мечту, лелеемую весь период между отпусками. Итак, хорошо проведенный отпуск — это цель, и ее достижение — значительный успех, его разумно трактовать как прибыль или выигрыш. Но за все хорошее приходится платить или расплачиваться, и, наверное, поэтому в теории принятия решений принято рассматривать потери, а прибыль или выигрыш — это отрицательные потери.

Вернемся к обсуждаемой ситуации. Во-первых, неприятности на дороге и аварии — это статистически устойчивые события, и можно предполагать наличие вполне определенных вероятностей попасть в аварию на любом из обсуждаемых видов транспорта. Но наших героев интересует не только вероятность аварии, но и потери, которые она за собой влечет.

При аварии самолета потери — это почти наверняка гибель, и поэтому потери весьма велики. Впрочем, потери возникают не только вследствие аварии. Есть и денежные потери — стоимость билета. Выигрыш от поездки самолетом — отрицательные потери — два лишних дня отпуска у моря.

При аварии поезда возникают потери вследствие повреждений и долгого пребывания в больнице, денежные потери — плата за проезд плюс еще потеря двух суток.

При поездке на автомобиле нашим героям грозят потери времени, сил и средств при подготовке автомобиля к

дальнему путешествию и, конечно, при аварии возникает не только возможность попасть в больницу или в «лучший мир», но и прямые потери времени, средств и сил при ремонте автомобиля. Наконец, следует учесть и стоимость поездки: она складывается из стоимости бензина, оплаты ночлега и, возможно, других различных расходов.

Наконец, при поездке на юг на мотоцикле, кроме реальных возможностей побиться в любой аварии и следующих за ней физических и материальных потерь, возникают потери вследствие утомительности многочасовой и многодневной поездки. Конечно, следует учесть и стоимость горючего, ночлега и разные непредвиденные траты.

Как видите, часть потерь выражается числами — в часах или рублях; другую часть трудно оценить в числах — едва ли вы сможете назначить цену за перелом руки, таза, сустава или за сотрясение мозга, а тем более за потерю собственной жизни. Выигрыш — отрицательные потери — здесь тоже есть: самоутверждение за счет преодоления трудностей, удобство собственного транспорта на отдыхе, ну и другие моральные факторы, о которых выше шла речь. Теперь для выбора наилучшего решения нужно сопоставить суммарные потери.

Решить нашу задачу непросто: нужно прежде всего суметь все потери сравнить между собой, для чего удобно их выразить в одних и тех же единицах, то есть перевести, скажем, в рубли теряемое время, моральный ущерб и возможные физические травмы (кое-что можно предложить и на этом пути, но мне не хочется отвлекаться). Поэтому для демонстрации метода рассуждений приведу значительно более простую задачу.

Школьник едет за город на автобусе. Билет стоит 50 коп., и мальчик-сладкоежка подсчитывает, сколько смог бы он купить мороженого на эту сумму. Образ мороженого волнует и завлекает, и школьник решает ехать без билета. Конечно, безбилетника может обнаружить контролер, и тогда ему грозит штраф и неприятности дома, но детский рационализм приводит к вечной борьбе с элементарными правилами, и он действует по принципу «риск — благородное дело».

На самом деле в риске никакого благородства нет, и сейчас мы обсудим ситуацию со школьником.

Возможны два исхода эксперимента: ω_0 — контролер есть, ω_1 — нет контролера, причем $P(\omega_0)=p$ — вероятность

появления контролера, $P(\omega_1)=1-p=q$ — вероятность непоявления контролера.

Школьник может принять два решения: d_0 — покупает билет, d_1 — едет без билета.

Возможные потери школьника: r_0 — стоимость билета, r_1 — величина штрафа за безбилетный проезд.

Теперь разумно подсчитать потери школьника в разных ситуациях.

Обозначим в соответствии с традицией через $L(\omega_i, d_j)$ потери в случае, когда принято решение d_j, а наступает исход ω_i. Так, $L(\omega_0, d_0)$ — это потери в случае, когда школьник купил билет и затем в автобусе появился контролер, так что $L(\omega_0, d_0)=r_0$ — стоимости билета.

Но когда речь идет о риске, то нужно учесть не просто те или иные потери или перечислить все их возможные варианты, а следует оценить какие-то средние потери при различных возможных решениях. В самом деле, когда вы переходите улицу с насыщенным автомобильным движением, то у вас есть шансы стать участником дорожно-транспортного происшествия. Однако вы переходите улицу, поскольку в среднем ваши потери не будут велики. Дети чаще попадают под автомобиль вследствие неправильной оценки именно этих средних потерь: ведь никто из них не хочет быть задавленным, но детям еще трудно оценить размеры возможного бедствия и сопоставить с шансами его осуществления.

Наиболее естественным мерилом обсуждаемых средних потерь служит, конечно, их математическое ожидание при выбранном решении d, так что нас интересует величина

$$\rho(d)=\mathsf{M}L(\omega, d),$$

где M — знак математического ожидания.

Она и называется риском при принятии решения d. Вычислим риск при обоих возможных решениях школьника. Если он выбирает решение d_0, то есть покупает билет, то

$$\rho(d_0)=L(\omega_0, d_0)p+L(\omega_1, d_0)q=r_0p+r_0(1-p)=r_0.$$

Это и естественно: если билет куплен, то появляется ли контролер или нет, потери школьника — это стоимость билета.

Если же школьник едет без билета, то

$$\rho(d_1)=L(\omega_0, d_1)p+L(\omega_1, d_1)q=r_1\cdot p+o\cdot q=r_1p.$$

И теперь видно, что означают слова «большой риск» или «небольшой риск»: если риск при решении d_0 много меньше риска при решении d_1 (то есть в нашем случае, если $r_0 \ll r_1 p$), то выбирать решение d_1 не следует: шансы оказаться в невыгодном положении очень велики. Если же имеет место обратное соотношение между величинами рисков, то имеет смысл рисковать — ехать без билета.

Произведем численный прикидочный подсчет. Если вероятность появления контролера $p=0{,}05$, то есть контролер посещает в среднем каждый двадцатый автобус, а величина штрафа 3 рубля, то риск безбилетника

$$r_1 p = 300 \cdot 0{,}05 = 15 \text{ коп.}$$

При цене билета $r_0=50$ коп. видно, что имеет место вторая ситуация, обеспечивающая безбилетнику в среднем заметный выигрыш.

В этом подсчете не учтена, конечно, моральная сторона дела. Но я и не призываю вас к нарушениям правил; скорее подобное вычисление — рекомендация о необходимой частоте контроля или размере штрафа для организации, которая должна поставить в невыгодное положение безбилетников. Впрочем, эта организация должна будет в вычислениях учесть и стоимость самого контроля, которая, естественно, возрастает с ростом вероятности p появления контролера — нужно иметь больше контролеров.

Вместо увеличения числа контролеров можно увеличить штраф. Скажем, при штрафе за безбилетный проезд в $r_1=20$ руб. и тех же величинах стоимости билета $r_0=50$ коп. и вероятности контроля $p=0{,}05$ имеем

$$r_0 = 50 < r_1 p = 20 \cdot 10^2 \cdot 0{,}05 = 100 \text{ коп.}$$

и теперь уже риск безбилетника оказывается слишком высоким. Примерно так и поступили в Варшаве: при цене билета 1,5 злотых размер штрафа — 50 злотых, и при той же вероятности $p=0{,}05$ положение оказывается не в пользу безбилетника.

Теперь, надеюсь, выпукло видна бессодержательность высказывания «риск — благородное дело»: бездумный риск, когда нет сопоставления величины риска при разных решениях, не благородное дело, а благоглупость.

Усложним немного задачу о выборе решения школьником: пусть он имеет возможность поехать в нужное место не только на автобусе, но и на поезде. Здесь будут

иными и цена билета, и величина штрафа, и вероятность появления контролера. Теперь школьник может принять одно из четырех возможных решений: d_0 — едет на автобусе, покупает билет, d_1 — едет на автобусе без билета, d_2 — едет на поезде, покупает билет, d_3 — едет на поезде без билета.

Решение новой задачи не оказывается сложнее. Но следует подчеркнуть одну особенность: теперь вероятности исходов зависят от принятого решения. Именно, вероятность появления контролера при поездке на автобусе и на поезде разные: при поездке на автобусе она равна p, а при поездке на поезде равна p'.

Таким образом, в общей ситуации распределение вероятностей на множестве исходов эксперимента, а вместе с ним и риск зависят от принимаемого решения.

Если вы хотите минимизировать свой риск, то выбираете решение, которое доставляет риску наименьшее из возможных значений.

Как вы, наверное, заметили, борьба с безбилетниками иногда активизируется — объявляется кампания за «обилечивание» пассажиров (этот удивительный термин я видел собственными глазами в автобусе на Кавминводах). Во время кампании резко возрастает количество контролеров. Пусть для определенности в обычное время вероятность появления контролера в автобусе равна 0,05, а во время кампании за «обилечивание» эта вероятность становится равной 0,6, причем пассажиру заранее неизвестно, началась ли кампания или нет. Иначе говоря, пассажир лишь знает, что вероятность p появления контролера в автобусе может принимать два значения $p_1=0{,}05$ или $p_2=0{,}6$.

В этой ситуации риск потенциального безбилетника зависит от вероятностей p_1 и p_2, и при каждой из этих вероятностей может быть риск, так что $\rho_1=\rho(p_1)$ и $\rho_2=\rho(p_2)$. Это, конечно, неудобно, так как безбилетнику не известно, началась ли кампания за «обилечивание», и это осложняет ему выбор наилучшей стратегии поведения.

Однако наблюдательный пассажир знает частоту, с которой происходят кампании борьбы с безбилетниками, и он полагает, что всякий раз вероятность обычной автобусной жизни равна, скажем, g, а вероятность кампании за «обилечивание» есть $1-g$. Эти вероятности называют априорными: они известны или, скорее, предполагаются

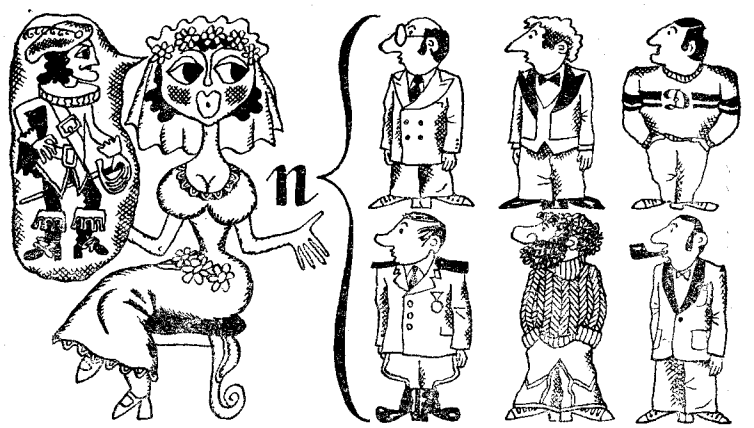
известными до проводимых наблюдений. Знание априорных вероятностей дает возможность вычислить математическое ожидание — риск, то есть взять риск, усредненный по априорным вероятностям.

Таким образом, средний риск для потенциального безбилетника, когда он едет в автобусе, будет равен

$$\bar{\rho}=g\rho_1+(1-g)\rho_2.$$

Риск, усредненный по априорным вероятностям, а вслед за ним и решение, доставляющее минимум среднего риска, называют байессовским.

Байессовский подход удобен — он ставит в соответствие каждому решению число, и это облегчает нахождение оптимального решения. Поэтому байессовский подход широко применяется в теории решений. Но здесь есть одно весьма уязвимое место: нужно знать априорные вероятности. Есть ситуации, когда априорные вероятности можно считать известными, например в задачах контроля качества продукции при установившейся технологии массового производства или в некоторых задачах медицинской диагностики. Но, к сожалению, есть много реальных ситуаций, где априорные вероятности не только неизвестны, но и вообще не имеют смысла. Скажем, в задачах обнаружения в системе ПВО не имеет смысла говорить об априорной вероятности появления самолета противника в данном районе. Поэтому при отсутствии априорных вероятностей или при значительных затруднениях с их заданием приходится отказываться от байессовского подхода.



Стратегия разборчивой невесты

Каждая девушка среди своих поклонников ищет Принца и, сопоставляя возможности с желаниями, старается «не промахнуться» — отправиться во Дворец бракосочетаний с наилучшим из возможных. Не легкая эта задача, иначе не было бы такого огромного количества разводов по инициативе молодых женщин.

Нет, я не собираюсь сейчас обсуждать различные аспекты столь сложной и вечно всех волнующей проблемы, но разберу одну ее простейшую математическую модель.

Пусть у красавицы имеется n женихов, и она с ними знакомится в случайном порядке. Ввести численные оценки «качества» жениха, конечно, трудно, особенно девушке, для которой главными характеристиками являются не только плохо формализуемые, вроде силы любви, элегантности и мужественности, но часто и такие, что «ни словом сказать, ни пером описать». Однако невеста, как бы она ни была капризна, в конце концов может решить, какой из любых двух женихов лучше, конечно, с ее невестинских позиций.

Мы разберем такую процедуру выбора жениха: невеста последовательно знакомится с женихами и после знакомства (период знакомства роли не играет) может отвергнуть его или назвать своим избранником. При этом предполагаются выполненными три важных условия:

во-первых, женихи появляются последовательно (одновременно нескольких сразу нет), во-вторых, вернуться к ранее отверженному жениху нельзя, наконец, в-третьих, количество женихов заранее фиксировано.

Возможно, вам представляются эти условия недостаточно реальными: в жизни иногда можно вернуться к ранее отверженному претенденту или вести небезызвестную игру сразу с несколькими женихами... Не буду спорить: на самом деле наша модель и не имеет в виду серьезные приложения в тонком деле сватовства. Поэтому приведу другие ситуации, математической моделью которых служит обсуждаемая нами модель.

Вы движетесь в автомобиле по трассе Москва — Симферополь и хотите пообедать. Вид дорожных ресторанов разнообразен: от павильонов «Пиво-Воды» до современных дворцов из бетона и стекла, да и меню, а тем более качество блюд далеко не одинаковы. Вот только шикарный внешний вид зданий не соответствует, вообще говоря, вкусовым качествам блюд. А вас почему-то интересует именно пища телесная... Вот снова ситуация, описываемая нашей моделью: вы последовательно подъезжаете к ресторанам и кафе, расположенным у дороги, и либо останавливаетесь и обедаете там, либо движетесь к следующему пункту в надежде на лучшее качество обеда. Вы, конечно, не поедете обратно — жаль времени. Количество точек, где можно пообедать, вам известно заранее: они обозначены на картах, издаваемых для автотуристов.

Впрочем, может быть вы настроены значительно более серьезно, и эта задача тоже вам не кажется достойной внимания... Поэтому я приведу производственную задачу, решение которой сводится все к той же математической модели.

Начальнику 1-го цеха поручили новую и весьма важную работу, и для ее выполнения предоставили возможность взять пятерых новых слесарей из тридцати только что присланных на завод после окончания производственно-технического училища. Начальник 1-го цеха, по договоренности, имеет преимущество — выбирает первым. Но если он отказался от очередного паренька, то слесаря тут же отдел кадров направляет в другой цех — потребности завода велики, и для 1-го цеха этот парень потерян безвозвратно. В то же время начальник 1-го цеха заинтересован в отборе пятерых лучших ребят. Какой стратегией выбора он должен воспользоваться? Посылаются в 1-й цех молодые рабочие поодиночке, и после собеседования, знакомства с документами и пробной работы начальник цеха должен решить, берет он рабочего или отсылает обратно и просит прислать другого. В математической модели еще предполагается, что начальник не только знает общее количество поступающих на завод (скажем, 30 человек), но и может относительно любых двух из них решить, какой лучше подойдет для выполнения порученной работы.

Иначе говоря, основным характеристическим свойством обсуждаемой задачи является возможность упорядочить объекты по качеству или, как часто говорят в теории решений, по предпочтению среди уже просмотренных объектов.

Задача выбора нескольких рабочих (m) из имеющихся (n) — это некоторое обобщение двух предыдущих, где производится выбор одного — наилучшего (жениха или ресторана), то есть ситуация, где $m=1$.

Прежде чем обсуждать результаты решения, приведу несколько более четкую постановку задачи. Для простоты пусть $m=1$.

Имеется n объектов, которые упорядочены по качеству. Можно представлять себе их качественную характеристику числом или точкой на вещественной оси: чем выше качество, тем больше число или тем правее расположена точка.

Знакомство с объектами происходит в совершенно случайном порядке, так что координата a_1 объекта, который появляется первым, может оказаться с равной вероятностью любой из имеющихся n точек. Аналогично второй объект с координатой a_2 может с равной вероятностью занять место любой из $n-1$ оставшихся точек. Таким образом, в результате последовательных наблюдений мы получим некоторый набор $a_{i_1}, a_{i_2}, \ldots, a_{i_n}$ координат, причем с равной вероятностью может появиться любая из $n!$ их возможных перестановок.

Точки (или объекты) появляются последовательно, а наша задача — остановиться, как только появится точка с координатой, самой большой из имеющихся, и, выбрав объект с этой координатой, прекратить наблюдения.

Но на самом деле мы же не знаем с абсолютной надежностью, что появился объект с самой большой координатой: мы-то можем его сравнивать лишь с уже наблюден-

ными, а не со всеми, и поэтому наверняка не ошибемся только тогда, когда лучший объект нам встретится на последнем шаге. Вся ситуация, конечно, укладывается в рамки теории вероятностей. Поэтому разумно так поставить задачу: нужно указать метод, приводящий к правильному решению с наибольшей вероятностью.

Какие же здесь имеются стратегии? Можно остановиться на первом же шаге, то есть выбрать точку с координатой a_1. Для разборчивой невесты это означает выбор первого же появившегося жениха. При такой стратегии она сразу же сможет надеть обручальное кольцо — предмет мечтаний всех девушек, но получает наилучшего из претендентов лишь с вероятностью $1/n$. Если претендентов много — n велико,— то при такой стратегии вероятность выйти замуж за наилучшего оказывается весьма маленькой.

Похоже, что при любой стратегии вероятность выбрать наилучшего жениха, или, в нашей более абстрактной постановке, остановиться на точке a_i с наибольшей координатой, будет всегда неограниченно уменьшаться с ростом числа n? Однако, к счастью, это не так.

Пусть значение n четно. Выберем такую стратегию: пропустим первые $n/2$ точек, а затем выберем первую точку, координата которой окажется больше всех предыдущих. Подсчет показывает, что вероятность остановиться на точке с наибольшей координатой при этой стратегии будет больше 0,25 независимо от величины n.

Итак, есть стратегия, приводящая к успеху с не такой уж малой вероятностью. Так как число n всех точек фиксировано, имеется и оптимальная стратегия, приводящая к успеху с наибольшей возможной вероятностью. Такая стратегия оптимального выбора состоит в следующем: пропускается некоторое определенное количество объектов s и затем выбирается первый объект, лучший всех предыдущих. Число s находится из двойного неравенства (n — заданное заранее число объектов)

$$\frac{1}{s+1}+\frac{1}{s+2}+\ldots+\frac{1}{n-1}\leqslant 1<\frac{1}{s}+\frac{1}{s+1}+\ldots+\frac{1}{n-1}.$$

При этой стратегии наилучший объект выбирается с вероятностью

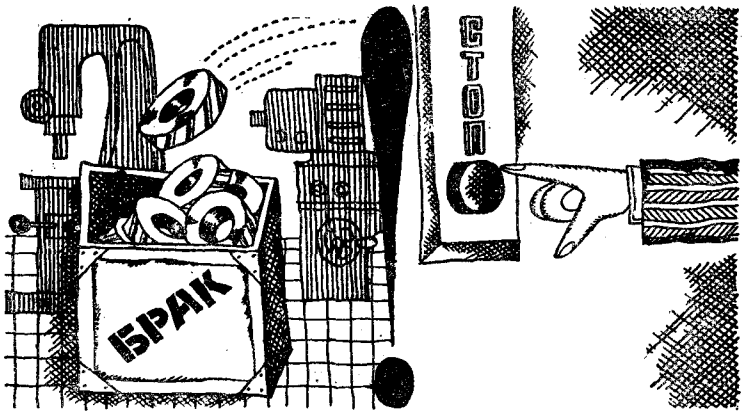
$$p_n=\frac{s}{n}\left(\frac{1}{s}+\frac{1}{s+1}+\ldots+\frac{1}{n-1}\right).$$

Например, вполне реально, что красавица имеет возможность выбора среди 10 женихов. Как вы легко можете проверить, подставив в предпоследнюю формулу $n=10$ и складывая дроби, в этом случае $s=3$, и, следовательно, оптимальная стратегия разборчивой невесты состоит в том, чтобы сначала пропустить первых трех женихов, а затем выбрать первого, который окажется лучше всех предыдущих. Здесь вероятность выбрать наилучшего среди всех десяти $p_{10}\approx 0{,}4$.

Если же n очень велико ($n\to\infty$), то при той же оптимальной стратегии вероятность выбрать наилучшего из всех возможных

$$p_n\approx\frac{1}{e}\approx 0{,}37$$

(e — основание натуральных логарифмов).



Управление качеством

Многие считают отцом автоматического регулирования Джеймса Уатта, создателя центробежного регулятора числа оборотов вала паровой машины. Но создание действующей паровой машины Уатту далось не легко. Американец Уильям Пиз*) пишет: «Первым станком в современном значении этого слова было устройство для расточки цилиндров, изобретенное Джоном Уилкинсоном в 1774 г. Уилкинсон не пользуется такой известностью, как Уатт, хотя именно его изобретение дало Уатту возможность построить действующую паровую машину. В течение десяти лет Уатт тщетно пытался изготовить цилиндр с нужной ему точностью. После одной из попыток он в отчаянии заявил, что в его цилиндре диаметром 45,7 см «в самом плохом месте отклонение от цилиндричности составляет приблизительно 9,5 мм». Однако уже в 1776 г. помощник Уатта Мэтью Болтон писал: «Мистер Уилкинсон расточил для нас несколько цилиндров почти без ошибки, отклонения в пятидесятидюймовом цилиндре, установленном нами

*) «Автоматическое управление», сборник статей, перевод с английского под редакцией В. Солодовникова, Изд-во Академия наук СССР, Москва, 1961 г. Здесь почему-то размеры даны в сантиметрах, а не в традиционных английских дюймах. Но, к сожалению, в этом сборнике переводов нет ссылок на оригиналы статей, и я привожу текст перевода.

в Типтоне, нигде не превышают толщины старого шиллинга»».

Я затрудняюсь теперь оценить относительную ошибку — не знаю толщину шиллинга, бывшего уже старым в конце XVIII века, но, по-видимому, точность увеличилась в несколько раз. Таким образом, станок для расточки цилиндров Уилкинсона сделал коммерчески выгодной паровую машину Уатта, и этот станок явился прямым предшественником современных точных металлорежущих станков.

Но какая же точность изготовления цилиндров и других металлических деталей нужна? Мне кажется, что важнейшим шагом в развитии промышленности была реализация принципа изготовления взаимозаменяемых деталей.

Посудите сами: разумно ли при производстве, скажем, автомобилей изготовлять каждую деталь отдельно, как уникальную, и затем подгонять их друг к другу. У нашего современника такое предложение вызовет ироническую усмешку. Но когда в 1789 г. Эли Уитни наладил производство мушкетов по заказу правительства США, реализовав свою идею изготовления мушкетов из взаимозаменяемых деталей, большинство специалистов того времени отнеслось к нему с недоверием и считало такую затею не имеющей практической ценности.

Сегодня же основой производства служат станки и другие машины, производящие высокоточные взаимозаменяемые детали, не только совершенно устраняя потребность в дорогостоящем ручном труде, но и осуществляя производство с точностью, значительно превышающей человеческие возможности. И все же не всегда качество продукции нас удовлетворяет, и проблема повышения качества продукции стоит очень остро. Поэтому я вновь хочу вернуться к этой проблеме, но обсудить с вами не приемочный контроль, о котором выше шла речь, а вопросы управления качеством продукции во время ее изготовления. Для такого управления необходим текущий контроль в самом процессе производства. Для определенности буду говорить о станке-автомате, изготовляющем стандартную продукцию, скажем болты, где контролируемым параметром служит длина болта.

Контроль состоит в измерении длины выбранных болтов и сопоставлении данных измерений с заданным размером, на который был настроен станок. Данные измерения, конечно, не совпадают точно с номиналом, всегда

имеются небольшие отклонения от среднего, вызванные болтанкой инструмента, колебаниями электроэнергии, неоднородностью материала и т. д. Эти отклонения — типичный пример случайной величины с непрерывным распределением вероятностей. Обозначим ее плотность вероятности через $p_0(x)$.

Однако когда нарушается работа станка — наступает разладка, то происходит либо заметное уклонение среднего значения измеряемого параметра от номинала, либо заметное увеличение разброса, либо и то, и другое. Иначе говоря, происходит изменение распределения вероятностей длин болтов; плотность вероятности нового распределения обозначим $p_1(x)$.

Итак, имеется две гипотезы: H_0 — станок работает нормально и H_1 — произошла разладка, причем нуль-гипотезе соответствует плотность $p_0(x)$, а альтернативной гипотезе $p_1(x)$.

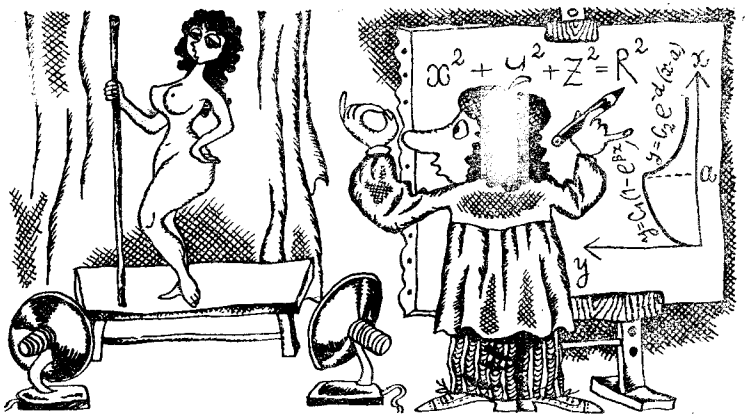
Произошла разладка — пошел брак, и, следовательно, задача управления качеством — как можно быстрее определить момент времени, когда наступила разладка, в этот момент остановить станок и обеспечить его подналадку.

Нам вновь, как и при выборе стратегии разборчивой невесты, нужно обеспечить оптимальную остановку. Однако здесь есть и существенная разница в постановках задачи.

Брак — негодные болты — это потери, причем их величина легко выражается в рублях. Переналадка станка влечет по крайней мере затраты рабочего времени — это тоже потери, выражаемые в рублях, а то и затраты материальных средств. Таким образом, остановка станка в любой момент приводит к некоторым потерям во вполне конкретных единицах — рублях, и можно ставить задачу оптимизации: выбирать момент остановки станка так, чтобы потери из-за разладки станка и его переналадки были минимальны. Ошибки первого и второго рода в данном случае — это пропуск бракованной продукции вследствие неправильно выбранного момента остановки и ложная тревога — остановка для подналадки, когда он работает нормально. Потери вследствие каждой из ошибок и средние потери легко вычисляются и выражаются в рублях.

В то же время, как мы уже обсуждали в предыдущем разделе, при поиске наилучшего жениха, ресторана или нового сотрудника весьма затруднительно задать численно величины потерь из-за неправильного выбора момента остановки (то есть ошибочного выбора номера или точки). Поэтому в предыдущем разделе и не ставилась задача оптимизации момента остановки, обеспечивающего минимум риска. Вместо этого мы ограничились стратегией, обеспечивающей выбор наилучшего объекта с наибольшей вероятностью.

Нахождение момента остановки в рассматриваемой задаче управления качеством, как и в задаче с разборчивой невестой, также достаточно элементарно и опирается на формулу Байесса.



Математическая модель

Детский рисунок — домик с трубой и дымом, круглое солнце с лучами, елочки и ромашки — модель окружающего мира, где ухвачены главные черты восприятия ребенка.

Художник — реалист или импрессионист — напишет этот же пейзаж по-разному: в зависимости от своей философской позиции художник подчеркнет в пейзаже разные его стороны, выразит в картине свое мировоззрение и настроение.

Но и ортодоксальный натуралист не сможет полностью, с абсолютной точностью воспроизвести натуру: даже если бы можно было воспроизвести все, что видит художник, то нельзя воссоздать на картине движение, запахи, звуки — жизнь во всем ее многообразии.

Обычно модель проще моделируемого объекта. Впрочем, скульптор называет моделью натурщика. Едва ли с общечеловеческих позиций статуя девушки с веслом сложнее самой девушки, но, возможно, эта терминология отражает соотношение замысла скульптора, обычно довольно глубокого, с внешностью натурщика, служащего лишь материалом для воплощения этого замысла.

Чертеж машины или схема радиоприемника—тоже модели, и не случайно при конструировании радиоаппаратуры создают последовательно блок-схему, принципиальную схему, монтажную схему. Все они моделируют будущий приемник не только с различной степенью детализации, но и отражают разные его стороны: если блок-схема описывает операции, осуществляемые различными блоками аппаратуры, то монтажная схема указывает, каким образом будут соединены элементы — сопротивления, конденсаторы, транзисторы — для выполнения нужных операций. Но когда приемник будет создан, его будут «доводить»: где-то придется убавить емкость или увеличить сопротивление. Монтажная схема — это модель, а не сам приемник.

В последние годы понятие «модель» используется широко и разнообразно в самых различных областях науки, техники, естествознания, в гуманитарных областях знания, в искусстве и даже художественной литературе. По-видимому, невозможно дать описание или тем более определение этому понятию, формально соответствующее всем его применениям и в то же время доступное специалистам столь разных областей. Да мне для дальнейшего это не так уж необходимо.

Я буду пользоваться значительно более ограничительным понятием математической модели — описанием изучаемого объекта на формальном языке, то есть с помощью чисел, уравнений различного вида (конечных, дифференциальных, интегральных, интегро-дифференциальных, операторных), неравенств или логических соотношений.

На карте шоссейных дорог, издаваемой для автотуристов, путь из Москвы до Харькова выглядит как ломаная линия, состоящая из прямолинейных кусков. Это, конечно, модель дороги. Можно построить и математическую модель — записать ломаную в виде серии линейных уравнений, «склеенных» в местах излома. Однако это очень грубая математическая модель дороги: здесь упущены многие важные для автотуриста детали от крутых поворотов, спусков и подъемов до названий деревень.

Прирост населения в городе пропорционален числу жителей. Такая математическая модель — линейное уравнение — верна лишь в довольно грубом первом приближении. Если же учесть количество стариков, детей, незамужних женщин, то модель прироста населения значительно усложняется. А если включить в модель такие факторы, как уровень образования, количество работающих женщин, уровень благосостояния и т. д., то математическая модель будет уже достаточно сложной, построить ее и изучить совсем не просто. Однако даже при учете подобных факторов модель может оказаться не очень близкой к

действительности: здесь не учтено множество случайных факторов — миграция населения, статистика браков и разводов и многое другое.

Обратимся к технологическому процессу производства бензина из сырой нефти. В процессе первичной переработки нефти бензин получается путем выпаривания: при подогреве нефти до определенной температуры более легкие бензиновые фракции переходят в газообразное состояние, и их отбирают в верхней части ректификационной колонны. Если считать температуру низа колонны и поступающего сырья фиксированными, то простейшей математической моделью, связывающей количество получаемого бензина с количеством поступающего сырья, будет линейное уравнение: если увеличить расход сырой нефти в 1,2 раза, то количество отбираемого бензина возрастет в 1,2 раза. Такова простейшая и очень грубая модель. Если учесть связь расхода сырья с температурой и давлением в колонне, то модель усложнится. Если же заметить, что через колону нельзя пропустить любое количество сырья, или неограниченно увеличивать температуру,— следует учесть технологические ограничения,—усложнение математического описания еще возрастет, однако и это не дает полную математическую модель процесса: на реальной установке первичной переработки нефти (АВТ—атмосферно-вакуумной трубчатке) измеряемых, регулируемых автоматически и управляемых параметров около двухсот, причем между многими из них есть сложные обратные связи. Даже если записать все подобные зависимости и ограничения, применять такую модель для управления нельзя будет даже при более высоком уровне вычислительной техники, чем сегодняшний. Но этого мало: есть еще плохо учитываемые случайные факторы — случайные изменения качественных характеристик поступающей на переработку нефти, случайные колебания температур, давлений, электроэнергии и тому подобное...

Не кажется ли вам проблема создания математической модели такого процесса безнадежной?

Еще более впечатляющей оказывается ситуация при попытке создать математическую модель живого. Скажем, можно ли составить математическую модель функционирования головного мозга собаки или человека, модель, учитывающую не только наличие нескольких миллиардов нервных клеток, но и связей между ними? Решение так поставленной задачи представляется совершенно безнадежным... Но если не учесть все многообразие процессов в мозге и построить модель, учитывающую лишь небольшую часть изменяемых величин и взаимосвязей, то едва ли такая модель нас устроит: она, по-видимому, не даст похожего, адекватного описания...

Как видите, термин «математическая модель» понимается довольно широко. Приведу высказывание из книги «Вероятностная модель языка» В. В. Налимова *), пишущего всегда остро и интересно.

«Часто совсем иной смысл вкладываем мы сейчас в термин «математическое моделирование», понимая под этим некоторое упрощение и весьма приближенное математическое описание сложной системы. Слово «модель» в этом случае противопоставляется закону науки, относительно которого предполагается, что он описывает явления природы некоторым «безусловным» образом. Одна и та же сложная система может описываться разными моделями, каждая из них отражает только какую-то сторону изучаемой системы. Это, если угодно, взгляд на сложную систему в некотором определенном и заведомо узком ракурсе. В этом случае, естественно, и не возникает задача дискриминации—различные модели могут иметь право на одновременное существование. Модель в этом понимании ведет себя в каком-то смысле так же, как описываемая ею система, а в каком-то другом смысле — иначе, ибо модель не идентична описываемой системе. Пользуясь лингвистической терминологией, мы должны были бы сказать, что математическая модель есть просто метафора».

Напомню, что в лингвистике под словом «метафора» понимают оборот речи, состоящий в употреблении слов и выражений в переносном смысле на основе какой-либо аналогии, сходства, сравнения, контраста. Зачем же нужно строить метафоры столь сложных систем, как первичная переработка нефти или мозг? Тем более, что построение таких математических моделей даром не дается: придется преодолевать трудности, подчас весьма значительные, и научного, и организованного, и психологического характера...

Я задал вам совсем не риторический вопрос, и его следует обсудить. Просто так, для удовольствия или увеличения списка научных работ строить математические мо-

*) Изд-во «Наука», 1974, стр. 136—137.

дели не нужно, даже, пожалуй, вредно. В то же время основной, если не единственный, способ познания — построение моделей, но не каких попало, а содержательных, дающих возможность выпукло увидеть какие-то интересные или нужные исследователю стороны изучаемого явления, объекта, процесса, погрузив в тень другие стороны. С иных позиций они могут оказаться более важными, и тогда нужно строить другую модель.

Все предрассудки от поверий дикарей до гаданья на кофейной гуще построены на ошибочном понимании причинно-следственных связей и неправильно построенных моделях, по которым осуществляются предсказания.

На каждом шагу мы строим какие-то (не математические) модели: то берем зонтик или надеваем легкое платье, посмотрев на небо, то, обнаружив отсутствие белил в соседнем магазине, закупаем их впрок, то, увидев на тротуаре одинокую собаку или пьяного, обходим их стороной.

Человек, попав в беду, тоже строит модель: девушка, увидев жениха, нежно беседующего с какой-то блондинкой, представляет себе измену (строит модель!) и, впадая в отчаяние, строит модель самоубийства или мести в зависимости от характера. Впрочем, вглядевшись, она узнает блодинку — жену его начальника, отчаяние заменяется надеждой, и тогда она строит уже иную модель ситуации и отправляется в парикмахерскую.

Словом, отчаяние и надежда связаны с моделью ситуации, которая нужна человеку для предсказания событий, а следовательно, и для выбора стратегии своего поведения. И известная детская игра «Что было бы, если бы...» это игра в предсказания; видимо, дети ее любят вследствие биологической необходимости научиться строить модели для правильного предсказания жизненных ситуаций.

Но вернемся к математическим моделям. Вычисление траекторий планет — построение математических моделей их движения — нужно было для предсказания поведения планет: моментов появления над горизонтом и захода, затмений и т. д.

Пока наука имела дело с довольно простыми системами, например с парой материальных точек в ньютоновской механике, можно было построить модели, верные всегда. Таковы законы Ньютона, Ома или Кирхгофа. Подобные законы носят характер некоторой абсолютной категории: закон может быть либо верен, либо ошибочен. И, конечно, он нужен для предсказания поведения тех величин (сил, токов, траекторий движения), взаимодействие между которыми трактует закон.

Впрочем, впоследствии выяснилось: ньютоновская механика, законы Ома или Кирхгофа применимы не всегда, и были указаны границы их применимости. Поэтому следует понимать, что абсолютность законов относится к данному уровню знаний, и на более высоком закон может быть пересмотрен.

В XX веке наука обратила пристальное внимание на системы значительно более сложные, чем получались в классической физике: здесь и сложные технологические процессы, и проблемы развитой экономики, и живая природа. Особенно широко начали заниматься сложными системами с развитием кибернетики, то есть с конца сороковых — начала пятидесятых годов нашего века. Сложные системы — их называют также диффузными, большими или плохо организованными — существенно отличаются от систем хорошо организованных. В хорошо организованных системах можно выделить процессы или явления одной физической природы, которые зависят от небольшого числа переменных, то есть системы с конечным, и притом небольшим числом степеней свободы. В то же время в системах сложных, плохо организованных, нельзя разграничить действие переменных различной физической природы. Например, в процессе первичной переработки нефти нельзя разделить действия материальных потоков, температур, давлений в различных частях огромной установки, состоящей из нескольких сорокаметровых ректификационных колонн с десятками, а то и сотнями тарелок, целых каскадов теплообменников и другой сложной аппаратуры.

Однако, несмотря на невозможность разграничить множество взаимосвязанных факторов, можно строить математические модели подобных процессов или объектов для решения задач прогнозирования их поведения и управления ими. Оказывается, нет необходимости включать в модель все переменные — подчас для решения четко поставленной задачи достаточно учитывать совсем небольшое их число. Но я не хочу забегать далеко вперед...

Теперь следует заметить: теория вероятностей — это математическая модель явлений природы, технических или социологических объектов или процессов, обладающих статистической устойчивостью. Именно об этом говорилось в разделе «Неопределенность и случайность». Но не грех подчеркнуть еще раз: теория вероятностей изучает случай-

ные события, а вместе с тем случайные величины, процессы или поля, но не изучает неопределенные события и связанные с ним процессы.

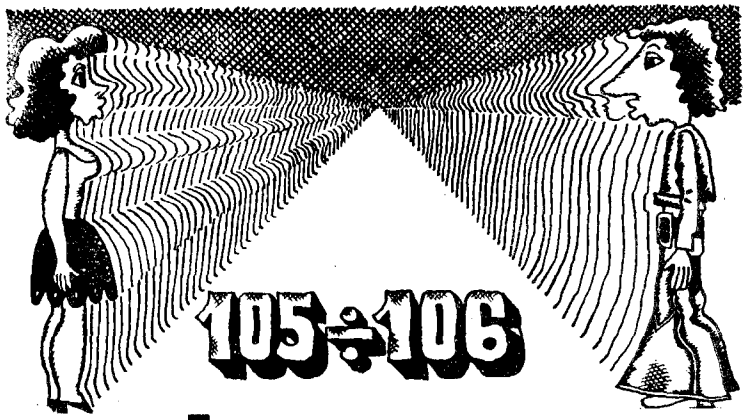
Само по себе распределение вероятностей — это тоже математическая модель. Я напомню вам одно из самых замечательных — распределение Бернулли или биномиальное распределение: если производятся независимые эксперименты и результат каждого из них может трактоваться либо как Успех, либо как Неудача, причем вероятность осуществления Успеха постоянна и равна p, то вероятность осуществления m Успехов при n экспериментах равна

$$P_n(m) = C_n^m p^m (1-p)^{n-m},$$

где $C_n^m = \frac{n(n-1)\dots(n-m+1)}{n!}$ — число сочетаний из n элементов по m.

Распределение Бернулли служит хорошей математической моделью таких явлений, как флуктуация плотности газа, количество вызовов на телефонной станции в течение некоторого интервала времени, дробового эффекта в вакууме, флуктуаций тока (в электронной лампе), флуктуаций интенсивности при сложении колебаний, наконец, одномерного случайного блуждания — простейшего варианта броуновского движения. Конечно, задача о количестве выпавших гербов при подбрасывании симметричной монеты — именно с этой задачи начинаются обычно традиционные курсы теории вероятностей — также содержится в математической модели. Но значение распределения Бернулли не только в возможности моделирования перечисленных или аналогичных задач; еще более существенно, что асимптотическое поведение этого распределения при одних ограничениях приводит к нормальному распределению, а при других — к распределению Пуассона, то есть к двум моделям, наиболее широко используемым в приложениях. Вам, конечно, знакомы эти распределения — о них говорится в любом курсе теории вероятностей. Напомню лишь, что нормальное распределение — математическая модель ошибок измерений, вариаций роста или веса животных одного вида и пола и многих подобных характеристик, встречающихся на каждом шагу, а распределение Пуассона — математическая модель радиоактивного распада, количества пожаров в городе, выпавших метеоритов или железнодорожных катастроф и многих других аналогичных процессов, часто называемых «редкими событиями».

Но хочу сейчас заметить: в литературе, посвященной применениям теории вероятностей, роль и универсальность этих распределений, как мне кажется, преувеличена. Когда читаешь книги, адресованные прикладникам, то подчас создается впечатление, будто в природе и технике никакие другие распределения просто не нужны — все описывается этими хорошо изученными распределениями или простейшими их функциями вроде логнормального распределения. Однако это ложное впечатление, и я в ближайших разделах приведу примеры, которые, надеюсь, покажут ограниченность упомянутых распределений как моделей в реальных задачах.



Вы, конечно, знаете эти слова из популярной песенки. Из песни слова не выкинешь, даже если это слово совсем не нужное, а то и ошибочное слово... Конечно, жаль девчонок, грустно подпирающих стенки на танцах или обнимающих подругу «за кавалера» (как здесь крепко устоялась мещанская терминология!) Но на танцы они ходят не только из желания потанцевать: среди кавалеров может попасться и долгожданный принц. Однако недостаток кавалеров объясняется не нехваткой парней — просто они предпочитают футбол или что-то еще: по данным нашей печати за 1972 год у нас в стране вплоть до 24-летнего возраста... на десять девчонок по статистике десять с половиной ребят, то есть на сто девиц приходится сто пять парней. Такое радующее девичью душу соотношение наблюдается не всегда.

Во время войн гибнут в основном мужчины, и в послевоенные годы соотношение между числом мужчин и женщин резко меняется и возникают ситуации, когда... на десять девчонок приходится всего пять ребят.

Речь сейчас пойдет о биологических закономерностях, о соотношении полов во всем животном мире и о механизмах, регулирующих это соотношение. Поэтому я буду придерживаться биологической терминологии.

Соотношением полов в популяции принято называть количество самцов, приходящихся на 100 самок. При этом различают соотношение полов в разные периоды: первичное соотношение полов — это соотношение зигот при оплодотворении; вторичное соотношение полов — при рождении и, наконец, третичное — соотношение полов в зрелой, способной размножаться популяции.

У различных животных в величине вторичного соотношения полов наблюдаются заметные отклонения в разные периоды, и в научной литературе в течение многих лет обсуждается вопрос о причинах, вызывающих эти колебания.

У человека вторичное соотношение полов, то есть количество родившихся мальчиков, приходящееся на сто родившихся девочек, в обычные периоды как раз и есть 105 ÷ 106. Таким образом, обычно вероятность рождения мальчика есть

$$105/(100+105)=0{,}512,$$

т. е. немного больше половины.

Но бывают в течение длительного времени значительные отклонения от этой величины, и статистическому анализу отклонений и поиску их причин посвящена большая литература. Скажем, статистически достоверно прослежено повышение вторичного соотношения полов, то есть относительный прирост рождающихся мальчиков во время и вскоре после длительных войн в странах, участвующих в войне. Так, в Германии в период первой мировой войны оно достигло 108,5, во время второй мировой войны в Великобритании и Франции вторичное соотношение полов возросло на 1,5—2%. На рис. 5 приведены данные о росте вторичного соотношения полов в Германии.

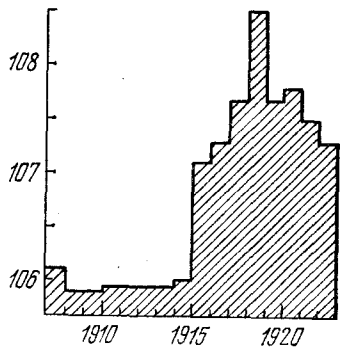


Рис. 5.

Для объяснения этого явления предлагалось много гипотез. Подмечено, что вторичное соотношение полов увеличивается с уменьшением возраста отца, во время войны вступают в брак более молодые люди, и ряд авторов

Таблица 3

Частота различных соотношений полов в семьях с 12 детьми (общее количество — около 5 миллионов рождений)

Количество		Наблюдаемое число случаев	Наблюдаемая частота на миллион	Ожидаемая частота на миллион (по биномиальному закону)	Знак разности: наблюдаемая частота — ожидаемая частота
сыновей	дочерей				
12	0	7	0,0007	0,0003	+
11	1	60	0,0056	0,0039	+
10	2	298	0,0279	0,0203	+
9	3	799	0,0747	0,0638	+
8	4	1398	0,1308	0,1353	—
7	5	2033	0,1902	0,2041	—
6	6	2360	0,2208	0,2244	—
5	7	1821	0,1703	0,1813	—
4	8	1198	0,1121	0,1068	+
3	9	521	0,0487	0,0448	+
2	10	160	0,0150	0,0127	+
1	11	29	0,0027	0,0022	+
0	12	6	0,0006	0,0002	+

считает это причиной увеличения вторичного соотношения полов. Другие связывают повышение вторичного соотношения полов во время и после войн с увеличением числа первородящих матерей, у которых доля мальчиков превалирует. Высказывались и другие гипотезы. Я не буду их обсуждать, но хочу отметить: ни одна из них не объясняет полностью наблюдаемое явление. И для его понимания — это же весьма сложная система — нужно придумать модель механизма явления, модель, которая могла бы объяснить способ, которым природа управляет воспроизведением потомства нужного популяции пола.

Первоначальные соображения о полной случайности и равновероятности рождения мальчика или девочки сразу отпадают вследствие обычно наблюдаемого вторичного соотношения 105 — 106 у людей или, скажем, 118 у собак. Ни у каких животных нет равного количества рождающихся самцов и самок.

Однако заранее все-таки не скажешь, кто же родится — мальчик или девочка, и поэтому естественной вероятностной моделью служит уже обсуждавшееся биномиальное распределение при отличной от $^1/_2$ вероятности рождения мальчика. Такая модель в первом приближении соответствует наблюдениям, но немного более глубокая

проверка приводит к выводу о ее неадекватности. Вот данные.

Исследовалось вторичное соотношение полов в семьях. Если бы биноминальное распределение было адекватной моделью рождаемости, то соответственно можно было бы предсказать частоты встречаемости семей с детьми одного пола или с заданным числом мальчиков и девочек. При обследовании и тщательной статистической обработке оказалось, что семьи с однополым потомством (одни мальчики или одни девочки) или с сильным преобладанием одного пола встречаются чаще, а семьи с равнополым соотношением или близким к нему встречаются реже, чем можно было бы ожидать, если принять, что соотношение полов — это только дело случая, и имеет место биномиальное распределение.

В таблице 3 приведены, как мне кажется, весьма впечатляющие данные. На рис. 6 представлены ожидаемая и наблюдаемая частоты.

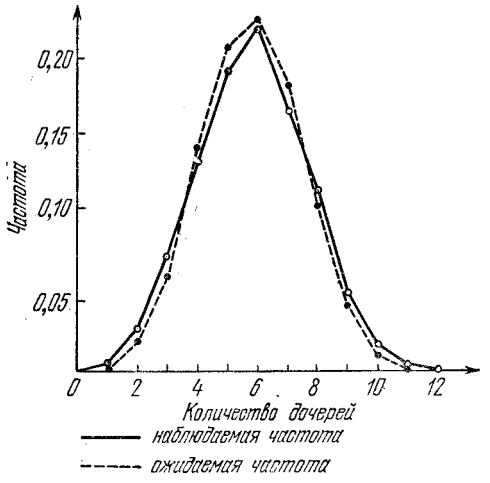


Рис. 6.

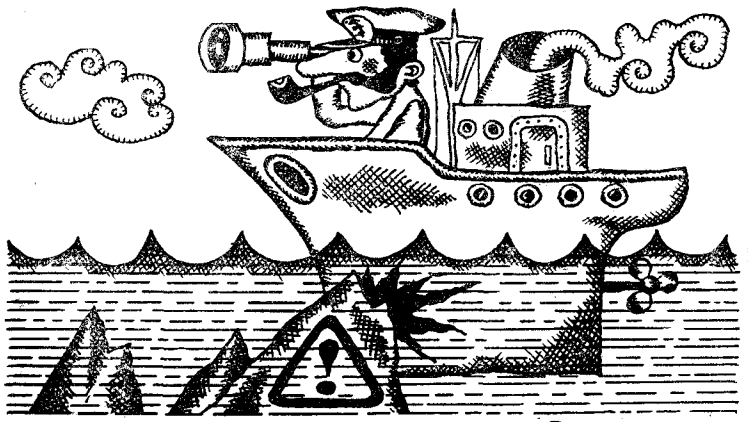
Колонка «Ожидаемая частота» подсчитана на основании биномиального распределения, где в качестве вероятности рождения мальчика взято относительное число мальчиков в этой серии наблюдений. Из сравнения ожидаемой частоты с наблюдаемой видно, что биномиальное распределение — модель здесь не очень хорошая, неадек-

ватная наблюдаемому распределению. Впрочем, в том же можно убедиться, применив, например, стандартные статистические критерии.

Эти данные и многие другие призывают к созданию модели, адекватной, соответствующей изучаемому явлению.

Такая модель предложена в 1965 году В. А. Геодакяном *). Он предполагает, что существует механизм обратной связи (отрицательной обратной связи, как говорят специалисты по теории колебаний и кибернетике), регулирующий вторичное соотношение полов при отклонениях третичного соотношения полов, то есть соотношения полов в зрелой популяции, и управляющим воздействием в замкнутой системе регулирования служат определенные гормональные факторы. Это, конечно, механизм статистический: меняются лишь вероятности рождения мальчика. Мне представляется модель В. А. Геодакяна вполне разумной, но, естественно, ее адекватность требует тщательной экспериментальной проверки. Таких данных в моем распоряжении нет, и поэтому я не буду более подробно обсуждать эту интересную тему.

*) О существовании обратной связи, регулирующей соотношение полов. Проблемы кибернетики, вып. 13, 1965.



Осторожно: задача свелась к линейной...

Задача свелась к линейной. Эта фраза в устах математика означает, что все основные трудности позади,— осталось только воспользоваться исчерпывающе разработанным аппаратом, и изучаемая проблема решена как с теоретической, так и с вычислительной точки зрения.

И все-таки... Если решалась прикладная задача, и линейная модель описывает некоторый реальный объект, то даже в рамках линейной задачи на пути исследователя могут встретиться различные сюрпризы. Об этом и пойдет дальше речь.

Начнем с простейшей, известной еще со школы системы двух алгебраических уравнений с двумя неизвестными:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2. \end{aligned}\right\} \qquad (1)$$

Ее решение имеет вид

$$x_1 = \frac{b_1a_{22} - b_2a_{12}}{a_{11}a_{22} - a_{12}a_{21}}, \quad x_2 = \frac{b_2a_{11} - b_2a_{21}}{a_{11}a_{22} - a_{12}a_{21}}. \qquad (2)$$

Алгебраическая теория систем линейных уравнений развита в предположении, что коэффициенты a_{ij} и правые части b_i известны точно — предположении, столь же естественном для математики, сколь неприемлемом в ее приложениях. В самом деле, когда система (1) является

математической моделью некоторого физического объекта, то коэффициенты обычно имеют вполне конкретный смысл. Они получаются в результате измерений или вычислений и часто известны весьма приближенно.

Сильно ли влияют погрешности в исходных данных на решение системы (1)? Специалисты по вычислительной математике уже давно обратили внимание на последствия от небольшого изменения коэффициентов системы (1) — иногда они оказываются катастрофическими. Вот классический пример: система

$$\left.\begin{aligned} x_1 + 10x_2 &= 11, \\ 10x_1 + 101x_2 &= 111 \end{aligned}\right\}$$

имеет решение $x_1 = 1$, $x_2 = 1$, а для системы

$$\left.\begin{aligned} x_1 + 10x_2 &= 11{,}1, \\ 10x_1 + 101x_2 &= 111 \end{aligned}\right\}$$

решением является $x_1 = 11{,}1$, $x_2 = 0$.

Подобные системы линейных алгебраических уравнений были названы плохо обусловленными, и первые рекомендации математиков были таковы: постарайся вовремя ее обнаружить и обойти, избегнуть в прикладной задаче. Но постепенно выяснилось: физические и технические задачи весьма часто сводятся к плохо обусловленным системам, и, игнорируя их, можно оказаться в положении того полковника, который шел в ногу, в то время как весь полк шагал не в ногу. Тогда интерес к плохо обусловленным системам возобновился на новой основе, и они были включены в общую проблему так называемых некорректных задач *).

При таком подходе всю теорию линейных алгебраических уравнений приходится начинать сначала с определения того, что следует понимать под решением системы. А затем, как обычно, требуется выяснить вопросы существования, единственности решения, найти способы его

*) Понятие корректности постановки краевой задачи для дифференциальных уравнений рассматривалось Ж. Адамаром еще в начале тридцатых годов, и им был построен пример ситуации, когда малое изменение начальных данных приводило к как угодно большому изменению решения. Впоследствии была понята важность понятия корректности задачи для физически осмысленных систем, и здесь следует упомянуть важные работы И. Г. Петровского.

построения и, конечно, убедиться, что вновь определенное решение не так бурно реагирует на небольшие изменения исходных данных — как говорят, является устойчивым.

Мы сейчас обратимся к системам линейных алгебраических уравнений, у которых коэффициенты a_{ij} и правые части b_i системы (1) (или ее естественного n-мерного аналога) являются случайными величинами.

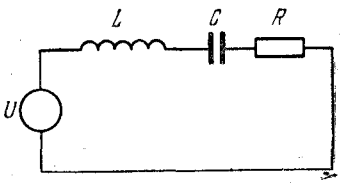


Рис. 7.

Но прежде чем перейти к их изучению, я хочу пояснить, насколько общую и широко распространенную модель мы здесь рассматриваем. Вы, конечно, понимаете, что большинство задач математического моделирования связано с изучением и решением уравнений конечных, дифференциальных, интегральных или более сложных. В реальных задачах параметры или коэффициенты уравнений определяются из эксперимента или назначаются экспертами. Для доведения решения задачи до числа, как правило, пользуются дискретизацией соответствующих уравнений, например переходят от дифференциальных к конечноразностным уравнениям. При использовании цифровых вычислительных машин такая дискретизация оказывается необходимой. В простейшем и в то же время наиболее распространенном случае дискретизация и упрощение уравнений приведут к системе линейных алгебраических уравнений.

Теперь я хочу пояснить, что трактовка коэффициентов и правых частей такой системы уравнений как случайных величин вполне естественна и разумна во многих прикладных задачах. Для этого приведу несколько примеров.

Рассмотрим последовательный электрический колебательный контур, состоящий из активного сопротивления R, емкости C и индуктивности L (рис. 7). Если величина входного напряжения (вещественного) есть U, то компоненты I_1 и I_2 комплексного мгновенного тока $I = I_1 + jI_2$ определяются, как известно, из системы линейных уравнений

$$\left.\begin{aligned} RI_1 - \left(\omega L - \frac{1}{\omega C}\right) I_2 &= U, \\ \left(\omega L - \frac{1}{\omega C}\right) I_1 + RI_2 &= 0. \end{aligned}\right\} \qquad (3)$$

Теперь, конечно, можно подставить коэффициенты этой системы в формулы (2) и получить решения I_1 и I_2. Но откуда взять величины R, L и C?

Если эта выборка из серийной партии, то точно значения величин R, L и C неизвестны. Мало того, ни в каком эксперименте эти величины не могут быть известны абсолютно точно, ибо, как уже говорилось выше, всегда любые измерения производятся с ограниченной точностью — таково фундаментальное принципиальное положение теории измерений. Неточности параметров R, L и C влекут за собой неточности в определении значений решения I, и решения уравнения (3) оказываются приближенными. При этом, как принято всюду в теории измерений, неточности при измерениях рассматриваются как случайные величины, а следовательно и решение I будет случайной величиной.

Другой пример относится к области экономического планирования. Пусть требуется увязать производство трех групп заводов как по линии взаимных связей, так и по линии точного выполнения заданных им планов на производство конечной продукции. Обозначим конечную продукцию соответственно через y_1, y_2, y_3, а валовые выпуски по группам заводов через x_1, x_2, x_3. Если a_{ij} — норма расхода продукта i-й группы заводов для изготовления одной тонны продукта на j-й группе заводов, то валовые выпуски и конечная продукция связаны следующей системой уравнений (i, j=1, 2, 3):

$$\left.\begin{aligned} (1-a_{11})x_1 - a_{12}x_2 - a_{13}x_3 &= y_1,\\ -a_{12}x_1 + (1-a_{22})x_2 - a_{23}x_3 &= y_2,\\ -a_{13}x_1 - a_{23}x_2 + (1-a_{33})x_3 &= y_3. \end{aligned}\right\} \qquad (4)$$

Однако нормы расхода a_{ij} — величины, конечно, усредненные, и задать их точно нельзя.

Обобщение этой задачи на n групп заводов очевидно: вместо системы (4) будет совершенно аналогичная система из n уравнений.

Подобных примеров можно привести сколько угодно. В других экономических, социологических, а то и технических задачах, также сводящихся к решению систем линейных алгебраических уравнений, коэффициенты a_{ij} подчас не могут быть вычислены, и тогда они задаются экспертными оценками. Вследствие очевидного субъективизма таких оценок величины a_{ij} также нельзя считать заданными точно.

Итак, если значения коэффициентов и правых частей получены в результате эксперимента или расчетов, проведенных с ограниченной точностью, то разумно считать их реализациями некоторых случайных величин. Тогда, скажем, система (1) содержит шесть случайных величин a_{11}, a_{12}, a_{21}, a_{22}, b_1 и b_2. Решением такой системы является, естественно, случайный вектор. Уточним: решением системы n линейных алгебраических уравнений со случайными коэффициентами и правыми частями будем называть случайный вектор $x = (x_1, x_2, \ldots, x_n)$ такой, что все случайные величины $\left(\sum_{k=1}^{n} a_{ik}x_k\right) - b_i$, $i = 1, \ldots, n$, с вероятностью 1 равны нулю.

Так как закон совместного распределения коэффициентов и правых частей заданной системы считается известным, то, говоря формально, можно вычислить n-мерный закон совместного распределения компонент вектора-решения x_1, x_2, ..., x_n, а следовательно, можно вычислить распределение каждой из компонент x_i. Однако провести вычисления довольно сложно. Например, линейная система десятого порядка содержит сто коэффициентов и десять правых частей, и для вычисления распределения каждой из компонент решения нужно вычислить стодесятикратный интеграл. Едва ли найдется много желающих взяться за такое дело... Но, главное, воспользоваться подобной формой решения еще сложнее! В самом деле, что должен делать физик, которому величина тока в колебательном контуре задана совместным распределением вещественной и мнимой частей этого тока? К счастью, при решении технических вопросов основную роль играет не само распределение, а некоторые его числовые характеристики: среднее значение, дисперсия, наиболее вероятное значение, размах и т. д. (заметим, что подобная ситуация является типичной во многих приложениях теории вероятностей и математической статистики).

Мы будем в дальнейшем характеризовать случайный вектор-решение алгебраической системы со случайными коэффициентами его математическим ожиданием, то есть вектором, полученным из заданного случайного заменой всех его компонент на их математические ожидания. Казалось бы, задача упростилась, но теперь возникли новые

трудности — введенное нами решение системы со случайными коэффициентами может вовсе не обладать математическим ожиданием и другими моментами! Следующие примеры показывают, что подобные опасения не лишены оснований. Вспомните, пожалуйста, закон Ома:

$$I=(1/R)U, \qquad (5)$$

где I — сила тока, U — напряжение, R — сопротивление. Написанное соотношение — это простейшее линейное алгебраическое уравнение, где $1/R$ — коэффициент.

Представим себе серийное производство радиоприемников. На одно и то же место в схеме ставятся «одинаковые» сопротивления, равные по номиналу 1000 Ом. Напряжение U в цепи здесь фиксировано, скажем, оно равно 100 В. Если взятое сопротивление точно соответствует номиналу, то сила тока в цепи будет равна 100/1000= =0,1А. Однако, как вы знаете, сопротивления, как и все другие заводские детали, изготовляются с ошибками, да и измеряются не без погрешностей. Поэтому на сопротивлениях, выпускаемых в продажу, пишут, например, 1 кОм±5%. Обычно в литературе еще со времен Гаусса для ошибок измерения, так же как и для ошибок изготовления, принимают нормальное распределение. Если сопротивление R, которое ставится в схему,— случайная величина, то и сила тока в цепи будет величиной случайной.

Какой же в среднем ток будет идти по цепи? Здесь следует понимать слова «в среднем» как среднее по всему множеству изготовленных схем. Так как случайная величина — сопротивление R, находится в знаменателе, то это среднее (математическое ожидание силы тока) не будет равно тому значению 0,1 А, которое получено при подстановке номинала в знаменатель. Но этого мало: в принятой модели искомое математическое ожидание просто не существует!

Для пояснения мне нужно написать интеграл, но, как уже говорилось, если вы мне верите на слово и не хотите разбираться в интегралах, пропустите их.

При нормальном распределении величины сопротивления R для математического ожидания тока I имеем

$$\mathsf{M}I = \frac{U}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{R} e^{-\frac{(R-R_0)^2}{2\sigma^2}} dR, \qquad (6)$$

где $R_0 = \mathsf{M}R$, равное в нашем примере 1000 Ом. Но этот

элементарный интеграл расходится вследствие особенности первого порядка при $R = 0$, и $\mathsf{M}I$ не существует!

Обратимся вновь к системе уравнений второго порядка (1) и ее решению (2). Рассмотрим простейшую ситуацию, когда $a_{11}=y$ является нормальной случайной величиной с математическим ожиданием a и дисперсией σ^2, а все остальные a_{ij}, так же как и b_i,— это какие-то постоянные величины. Тогда математическое ожидание x_1

$$\mathsf{M}x_1 = \int_{-\infty}^{\infty} \frac{b_1a_{22} - b_2a_{12}}{ya_{22} - a_{12}a_{21}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-a)^2}{2\sigma^2}} dy. \qquad (7)$$

Вследствие того, что знаменатель первого сомножителя подынтегрального выражения обращается в нуль при $y = a_{12}a_{21}/a_{22}$, причем это нуль первого порядка, а второй сомножитель в нуль нигде не обращается, то интеграл (7), такой же как и (6), обязательно расходится*), и, таким образом, в ситуации, когда один из коэффициентов случаен и распределен нормально, у компонент решения не существует математического ожидания!

Теперь, очевидно, что если не только a_{11}, но и все коэффициенты и правые части случайны и нормальны, то, вообще говоря, у компонент не существует ни математического ожидания, ни дисперсии, ибо всегда найдутся комбинации значений коэффициентов, при которых знаменатель обращается в нуль и интегралы будут расходиться. Обратимся к рассмотренному примеру с колебательным контуром, сводящемуся к системе (3). При изготовлении электрических сопротивлений, индуктивностей, емкостей всегда присутствуют погрешности, и величины R, L, C имеют разброс. В соответствии с упомянутыми выше положениями теории ошибок эти погрешности должны подчиняться нормальным распределениям. Проведенное выше рассуждение означает, что математическое ожидание и дисперсия у решения системы не существует, то есть средняя величина тока в контуре не существует, а его дисперсия, а вместе с ней мощность равны бесконечности! По-видимому, ни один инженер-электрик с подобным утверждением не согласится.

Случайные величины, не имеющие математического ожидания и дисперсии, как правило, не представляют

*) Полагаем, что числитель отличен от нуля; в противном случае следует рассмотреть $\mathsf{M}x_2$.

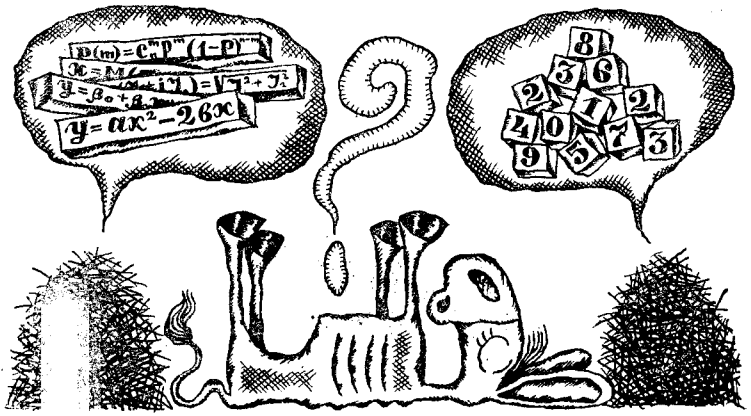
интереса в реальных задачах, и поэтому нам следует в математической модели исключить заранее ситуации, когда компоненты решения не имеют математического ожидания и дисперсии.

В связи с некоторыми задачами геофизики В. М. Глоговский и я столкнулись с этим кругом вопросов. К нашему удивлению, оказалось, что, когда для коэффициентов используются наиболее широко распространенные вероятностные модели, их простейшие числовые характеристики такие, как математическое ожидание и дисперсия, как правило, не существуют!

Однако величины, с которыми мы имели дело, вполне реальны — это скорости распространения упругих волн в земной коре, глубины залегания определенных горизонтов и т. д. И поэтому формальное отсутствие у них математического ожидания, очевидно, противоречит здравому смыслу и практике. Вывод один: тем хуже для модели, именно она приводит к таким несоответствиям, модель неадекватна, ее нужно изменить, заменить другой. Но при этом, конечно, полезно, а то и необходимо, понять причины несоответствия модели реальной ситуации.

Оказалось, что противоречия возникают вследствие учета хвостов распределений. И мы отказались в задачах решения систем линейных уравнений со случайными коэффициентами от рассмотрения распределений, подобных нормальному или экспоненциальному, с хвостами, уходящими в бесконечность, и обратились к распределениям, сосредоточенным на конечных интервалах, финитным, как говорят математики. Примерами финитных распределений служат равномерное распределение и так называемое усеченное нормальное распределение, то есть нормальное распределение, у которого исключены из рассмотрения все значения, лежащие вне некоторого интервала,— «обрезаны хвосты» распределения.

Для рассмотрения подобных финитных распределений в общем-то нет никаких принципиальных ограничений, но есть психологический барьер: вследствие причин, о которых я говорил, все верят в универсальность нормального распределения. Однако всякая универсальность имеет границы... И если модель нормального, экспоненциального или какого-нибудь другого распределения вам не подходит — мешают «хвосты», то следует, не жалея, «обрезать хвосты» и в качестве модельного брать финитное распределение.



Решение: формула или число?

Вы, конечно, знаете ответ: как когда... Формула, если она достаточно простая и наглядная, дает возможность увидеть качественную картину: как будет изменяться решение при изменении входящих в выражение параметров, как будет вести себя решение при очень больших или очень малых значениях переменных и т. д. Эти сведения подчас необходимы не только теоретику, но и инженеру-практику, естествоиспытателю, экономисту. Но все же рано или поздно нужно доводить решение до числа. И тогда, если формульное решение задачи сложное или оно представляется недоступным, возникает проблема вычислимости решения: нужно придумать метод вычисления приближенных значений решения или указать вычислительный процесс, который позволяет последовательно получать все более точные приближения к решению.

В обсуждавшейся в предыдущем разделе задаче определения математического ожидания и дисперсии компонент решения системы линейных алгебраических уравнений со случайными коэффициентами возникает как раз такая ситуация.

Принципиально задача определения первых и вторых моментов сводится к вычислениям соответствующих многократных интегралов. Однако написать формулы и получить числовые значения — это совсем разные задачи: при

численных квадратурах многократных интегралов подчас возникают колоссальные вычислительные трудности.

Поэтому перед нами возникает необходимость сделать вычислимой задачу получения числовых характеристик компонент решения системы.

Для этого в настоящее время используют два подхода: либо метод статистического моделирования (метод Монте-Карло), либо разложение решения в ряд Тейлора в окрестности математического ожидания коэффициентов и правых частей системы уравнений.

При использовании метода Монте-Карло следует по правилу Крамера записать формальное решение системы — это и есть формулы (2) предыдущего раздела, и затем, в соответствии с распределениями коэффициентов и правых частей, по таблице случайных чисел выбирать их значения, вычислять детерминанты, входящие в формулы Крамера, и таким образом получать реализации компонент вектора решения. Затем, набрав достаточное количество реализаций, то есть получив выборку достаточно большого объема, следует взять эмпирическое среднее — среднюю арифметическую выборки.

Однако если система не слишком низкого порядка, то метод Монте-Карло здесь приводит к огромному объему вычислений и сравнительно не высокой точности полученных приближений.

Второй метод таит в себе довольно неожиданные «подводные камни». Обратимся вновь к примеру с электрическим контуром, то есть к системе (3) предыдущего раздела.

Предположим для определенности, что R, L и U — постоянные величины, принимающие значения $R = 1$ Ом, $L = 10^{-3}$ Гн, $U = 5$ В, в то время как емкость C — случайная величина, равномерно распределенная в интервале $\gamma = (10^{-9} - 0{,}01 \cdot 10^{-9},\ 10^{-9} + 0{,}01 \cdot 10^{-9})$, иначе говоря, в интервале 10^{-9} Ф $\pm 1\%$. Тогда модуль комплексного тока $|I|$ будет вполне определенной функцией от переменной C — вы ее можете без труда получить. Для определения математического ожидания модуля тока на резонансной частоте $\omega = 1/\sqrt{L \cdot MC}$ следует вычислить интеграл от модуля тока по интервалу γ (поделив интеграл на длину интервала), и в результате вычислений получится величина $MI = 1{,}5$ А.

Если же, как это рекомендуется практическими руководствами, разложить выражение для $I = I(C)$ в ряд Тейлора до первых двух членов и вычислить теперь математическое ожидание, то получится 5А.

Таким образом, ошибка в величине емкости, равная 1%, приводит к погрешности в искомой величине MI, достигающей 330%.

Отметим, что использование для нахождения математического ожидания решения первых двух членов разложения в ряд Тейлора эквивалентно следующему приему: коэффициенты системы предыдущего раздела заменяются их математическими ожиданиями, и решается новая система уравнений с детерминированными коэффициентами. Этот путь вообще кажется привлекательным: похоже, что, заменив коэффициенты их математическими ожиданиями и решив получившуюся одну систему алгебраических уравнений, можно вычислить удовлетворительную оценку математического ожидания решения.

Приведу еще один пример, показывающий крайнюю грубость, а потому и непригодность такого подхода без предварительной оценки возможной погрешности.

Рассмотрим задачу вычисления экстремума параболы

$$y = ax^2 - 2bx, \qquad (*)$$

у которой коэффициенты a и b — независимые случайные величины, распределенные равномерно в интервалах соответственно $(10^{-3}, 1)$ и $(5,7)$.

Стационарная точка x_0 и величина экстремума y_0 здесь также случайные величины:

$$x_0 = \frac{b}{a}, \quad y_0 = -\frac{b^2}{a}. \qquad (**)$$

Прямое вычисление математических ожиданий этих случайных величин с учетом их распределений дает значения

$$Mx_0 = 87{,}3, \quad My_0 = -251{,}1.$$

Если же заменить в уравнениях (*) и (**) случайные величины a и b на их математические ожидания $Ma = (1-10^{-3})/2$, $Mb = 6$ и вычислить теперь координату $\hat{x}_0$ стационарной точки x_0 и величину экстремума $\hat{y}_0$, то получаются величины

$$\hat{x}_0 \approx 12, \quad \hat{y}_0 = -72.$$

Как видите, ошибка получается недопустимо большой, и на самом деле может быть как угодно большой при

«неудачном» распределении коэффициента a. Например, если a распределено равномерно в интервале $(-0{,}5,\ 1)$, то $\mathsf{M}a = 0{,}25$ и оценки $\hat{x}_0$, $\hat{y}_0$ принимают конечные значения, в то время как в действительности математические ожидания $\mathsf{M}x_0$, $\mathsf{M}y_0$ не существуют, ибо $\int_{-0,5}^{1} \frac{da}{a}$ расходится.

Приведенные примеры и многие другие практически интересные задачи показывают, что подобные способы вычисления математического ожидания решения системы линейных алгебраических уравнений со случайными коэффициентами могут приводить к очень грубым оценкам и прямым ошибкам.

Это предостережение особенно нужно иметь в виду, когда коэффициенты системы зависимы. Вот парадоксальный пример. Пусть матрица коэффициентов системы (1) предыдущего раздела имеет вид

$$A = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix},$$

где α — случайная величина, равномерно распределенная на интервале $[0,\ 2\pi]$. При любой реализации случайной величины α $\det A = 1$ и поэтому решение уравнения (1) предыдущего раздела, в котором коэффициенты заменены соответственно элементами матрицы A, существует для любого вектора b, причем $\mathsf{M}x_1 = \mathsf{M}x_2 = 0$. Вместе с тем математические ожидания всех элементов матрицы A равны нулю, и следовательно, если заменить коэффициенты рассматриваемой системы на их математические ожидания, то получающиеся при этом уравнения вообще не имеют решения, а значит, и математического ожидания.

Заметим, что этот пример не является искусственным, ибо матрица A — это матрица поворота ортогональной системы координат на угол α.

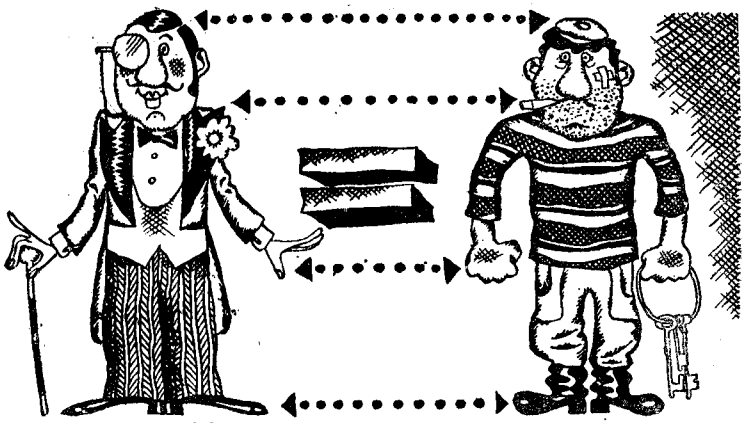
Можно ожидать, что точность определения математического ожидания решения системы повысится, если увеличить число членов в формуле Тейлора. Так оно, конечно, и есть. Однако нахождение следующих членов разложения требует преодоления серьезных вычислительных трудностей, ибо число слагаемых растет очень быстро (ведь речь идет о ряде для функций многих переменных!), а их вид зависит от конкретно предложенной задачи. Кроме того, простое увеличение числа членов ряда

без оценки точности вычислений математического ожидания может привести к абсурду.

Например, если при нахождении математического ожидания модуля тока в колебательном контуре взять не два члена формулы Тейлора (как мы сделали выше), а три, то получится оценка $\mathsf{M}I = -83{,}3$, бессмысленная совершенно, ибо модуль тока I, а вместе с тем и его математическое ожидание положительны.

Вместе с тем представление решения системы (1) предыдущего раздела в виде ряда Тейлора обладает одной очень важной особенностью. В отличие от решения в форме Крамера, являющегося дробно-рациональной функцией многих переменных — коэффициентов, отрезок ряда Тейлора — многочлен от этих переменных. Поэтому, если отрезок степенного ряда достаточно хорошо представляет решение, то само вычисление математического ожидания (т. е. нахождение интегралов от алгебраического многочлена) не вызывает никаких затруднений. Следовательно, разумно отказаться не от представления решения в виде степенного ряда, а лишь от использования ряда Тейлора. При этом от нового ряда мы хотим потребовать, чтобы он, во-первых, достаточно быстро сходился и, во-вторых, чтобы его члены достаточно просто, желательно единообразно, получались из коэффициентов заданной системы уравнений и, в-третьих, чтобы оценка точности остаточного члена была вычислимой. Всем этим требованиям удается удовлетворить, если использовать для приближенного решения системы итерационные методы.

Не буду приводить здесь разработанный нами алгорифм для вычисления моментов решения — речь в этой книжке идет о другом. Я рассказал о проблеме решения систем линейных алгебраических уравнений со случайными коэффициентами для демонстрации того, как существенно зависит решение задачи от выбранной модели, в частности, в этой задаче — от типа распределения. Если рассматривать вероятностную модель случайных коэффициентов в виде нормального распределения или ему аналогичного с хвостами, уходящими в бесконечность, то осмысленное решение не получается. Если же «обрезать хвосты» и рассматривать модели коэффициентов в виде финитных распределений, то получаются разумные результаты. Поэтому, прежде чем решать задачу, следует уделить внимание обдумыванию выбираемой математической модели.



Идентификация преступников — бертильонаж

Вы с волнением следите за действиями детектива, конечно, удобно расположившись на диване, и вот подозреваемый в убийстве выслежен и задержан. Вы откладываете книжку и приятно потягиваетесь. Однако если детектив настоящий, то его работа на этом далеко не окончена: ему надо установить личность задержанного, доказать, что, во-первых, именно он убил прохожего в подъезде и, во-вторых, предъявивший паспорт на имя Пантюхина А. Ф. на самом деле бандит-рецедивист, известный под кличкой «Море».

Непростая задача идентификации, то есть установления идентичности, одинаковости двух лиц. Сто лет назад в криминальной полиции некоторых стран Европы сопоставляли фотографии (анфас и профиль) и кое-какие описательные характеристики, но попробуйте перебрать десятки тысяч фотографий и выбрать похожие! Поэтому проблема идентификации преступников стояла очень остро.

Альфонс Бертильон занял в 1879 году пост писаря полицейской префектуры Парижа, и его деятельность состояла в заполнении карточек с описанием личности преступников. Эти записи носили весьма расплывчатый характер: высокий, среднего роста или низкий, лицо со шрамом или обычное, а то и просто «никаких особых примет».

Бартильон вырос в семье, интересующейся естествознанием,— его отец был уважаемым врачом, статистиком и вице-президентом Антропологического общества Парижа. Альфонс знал о Дарвине, Пастере, Дальтоне, Гей-Люссаке и, наконец, много раз слышал имя Адольфа Кетле — бельгийского математика и статистика, не только добившегося успехов на математическом поприще, но и стремившегося доказать, что строение человеческого тела подчинено определенным закономерностям. Дальше я выборочно цитирую книгу Ю. Торвальда «Сто лет криминалистики» *), написанную далеко не банально: это настоящее научное сочинение, где детективные истории служат иллюстрациями к научным достижениям криминалистов, и поэтому книга читается с захватывающим интересом.

— «И вот в июле 1879 года, когда Бертильон, изнемогая от парижской жары, сидел и до одурения заполнял трехтысячную или четырехтысячную карточку, его вдруг осенила идея, которая родилась, как он позднее признавался, от сознания абсолютной бессмысленности его работы и одновременно из воспоминаний детства. Почему, спрашивал он себя, напрасно тратятся время, деньги и усилия людей на установление тождества уголовников? Почему цепляются за старые, грубые, несовершенные методы, когда естествознание установило возможности безошибочно отличать одного человека от другого по размерам тела?»

«Бертильон вызвал удивление и насмешки других писарей, когда в конце июля стал сравнивать фотографии арестантов. Он сравнивал форму ушей и носов. Громкий смех вызвала просьба Бертильона разрешить ему обмерять регистрируемых заключенных. Ко всеобщей потехе ему это позволили. С мрачным ожесточением он за несколько недель обмерил довольно большое число арестованных. Измеряя их рост, длину и объем головы, длину рук, пальцев, стоп, он убедился, что размеры отдельных частей тела разных лиц могут совпадать, но размеры четырех, пяти частей тела одновременно никогда не будут одинаковы.

Душная жара августа вызвала приступы мигрени и носовые кровотечения, но Бертильона, казалось бы никчемного и бесцельного, захватила «власть идеи». В середине августа он написал доклад, в котором изложил, как можно

*) Изд-во «Прогресс», 1975.

безошибочно идентифицировать преступников. Этот доклад он направил Луи Андрие, занимавшему с марта 1879 года пост префекта полиции Парижа, но не получил никакого ответа.

Бертильон продолжал работать. Каждое утро до начала работы он посещал тюрьму Ла Сантэ. Там тоже насмехались над ним, хотя и разрешали производить измерения. Когда первого октября его повысили в должности, он передал префекту второй доклад, в котором, ссылаясь на закон Кетле, отмечал, что размеры костей взрослого человека всю его жизнь остаются неизменными. Он утверждал: если вероятность совпадения роста людей представляет собой соотношение 4:1, то рост плюс еще одно измерение, например длина тела до пояса, уменьшает вероятность совпадения до 16:1. Если же сделать 11 измерений и зафиксировать их в карточке уголовника, то по исчислениям вероятности шанс найти еще одного уголовника с такими же данными составит 4 191 304:1. Если оперировать четырнадцатью измерениями, то этот шанс снизится до соотношения 286 435 465:1. Набор частей тела, которые можно подвергнуть измерению, очень велик: кроме роста человека, можно измерить длину и ширину его головы, длину пальцев, предплечья, стоп и так далее. «Все имеющиеся способы идентификации поверхностны, ненадежны, несовершенны, порождают ошибки»,— писал он. Его же метод вселяет абсолютную уверенность и исключает ошибки. Кроме того, он, Бертильон, разработал систему регистрации карточек с данными измерений преступников, благодаря которой за несколько минут можно установить, имеются ли данные арестованного в картотеке».

Таким образом, Бертильон для идентификации преступников предложил использовать некоторый набор антропометрических измерений.

Конечно, потребовалось немало времени и душевных сил для преодоления косности и недоверия. Но добился успеха и признания Бертильон, как обычно, вследствие стечения нескольких благоприятных обстоятельств, когда выдуманная им система регистрации и огромная проведенная работа дали возможность идентифицировать личности нескольких крупных преступников.

Система Бертильона — бертильонаж состояла в измерении одиннадцати величин: роста, размаха рук, ширины груди, длины и ширины головы, длин левой стопы, среднего пальца левой руки, левого уха и еще чего-то.

Теперь мне следует прокомментировать бертильонаж. С наших позиций речь идет о построении математической модели человека в виде набора чисел $(x_1, x_2, ..., x_n)$, то есть в виде точки в n-мерном пространстве или n-мерного вектора. При этом бертильонаж — это использование одиннадцатимерного пространства.

Бертильон использовал подсчет шансов встретить двух людей с одинаковыми значениями измеряемых размеров тела. В книге Ю. Торвальда утверждение «...нет на земле двух человек, у которых совпадали бы размеры отдельных частей тела и что шанс встретить двух совершенно одинаковых по росту людей расценивается как один к четырем» приписывается Кетле и написано, будто отец и дед Бертильона (последний был математиком и естествоиспытателем) проверяли утверждение Кетле.

Мне представляется, что в этих подсчетах допущены по крайней мере две ошибки. Во-первых, вероятность совпадения роста двух случайно выбранных людей не равна 1/4, она раза в 3—4 меньше. Во-вторых, в приведенном подсчете перемножаются вероятности совпадений размеров выбранных частей тела, то есть предполагается статистическая независимость, скажем, ширины и длины головы или роста и размаха рук. Но статистическая независимость здесь не имеет места — эти величины сильно корреллированы, и поэтому мы удивляемся, встретив высокого человека с маленькой головой или низкорослого с длинными руками, достигающими почти до пола.

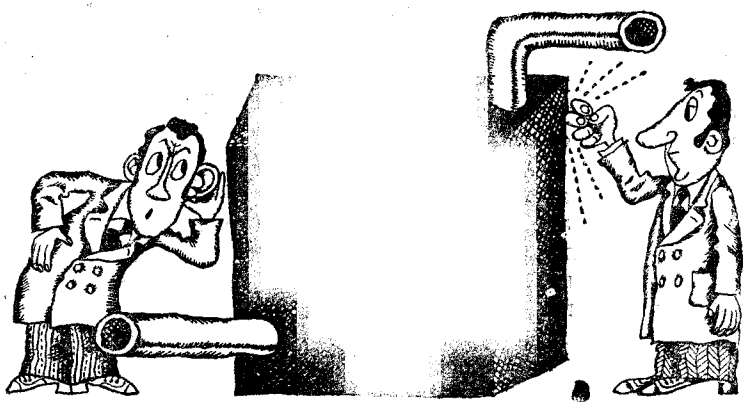
Допущенные в подсчете ошибки в некотором смысле компенсируют друг друга, и шансы обнаружить двух людей с одинаковыми значениями размеров всех одиннадцати измеряемых Бертильоном величин действительно очень малы — вот основа успеха бертильонажа.

Бертильонаж на некоторое время завоевал признание, но его повсеместному использованию мешал ряд обстоятельств, и главное из них — сложность реализации. Для проведения измерений размеров частей тела нужно, чтобы обследуемый этому не сопротивлялся,— преступник должен сидеть спокойненько и подставлять то голову, то руку, то стопу... Проводящий же измерения должен делать свое дело тщательно и аккуратно. Уровень культуры полицейских и тюремщиков был достаточно низок, и данные проведенных ими измерений не вызывали доверия. Бертильонаж, хотя и завоевал в некоторых странах приверженцев, не был принят повсеместно.

Подобная ситуация довольно часто встречается: безупречная теоретическая разработка не внедряется в производство или не используется в естествознании вследствие сложности реализации. Вот здесь и необходим здравый смысл: нужно либо самому придумать простой способ реализации, либо искать людей, способных это придумать, а не сетовать на «непонимание» практиков. В одной из лабораторий нашего института я увидел плакат такого содержания: «Если хочешь добиться — ищешь возможности, если не хочешь — ищешь причины». С этим нельзя не согласиться.

Бертильон не смог упростить настолько систему, чтобы снятие данных не требовало ни квалификации обслуживающего персонала, ни особой его добросовестности.

Поэтому через некоторое время на смену бертильонажу пришла дактилоскопия — сопоставление отпечатков пальцев. Ее история тоже весьма поучительна, но она имеет отношение к другой теме — распознаванию образов,— которую здесь я не буду затрагивать.



Идентификация технических объектов

При разработке нефтяных месторождений методами законтурного и внутриконтурного заводнения, которые широко используются на отечественных нефтепромыслах, нефть из скважины на поверхность поступает вместе со значительным количеством воды. При грубой оценке последние годы в среднем по стране добывается столько же воды, сколько и нефти.

Вода, идущая вместе с нефтью из нефтеносного пласта — пластовая вода, очень вредная: она сильно минерализована, содержит до нескольких тысяч миллиграммов солей в литре жидкости.

Как вы, конечно, знаете, нефть легче воды, и на нефтепромыслах отделяют основную часть воды от нефти путем отстоя в резервуарах. Затем отстоявшуюся смесь разделяют: верхнюю часть — нефть перекачивают в нефтепроводы, а нижнюю — воду закачивают через нагнетательные скважины обратно под Землю. Однако полностью так избавиться от воды и содержащихся в ней солей нельзя, и некоторое количество этой вредоносной минерализованной воды (порядка нескольких процентов или долей процента от объема нефти) транспортируется вместе с нефтью на нефтеперерабатывающие заводы. Наличие такого большого количества солей приводит к быстрой коррозии аппаратуры, и если не организовать извлечение солей из нефти до ее поступления на переработку, то установки

будут все время выходить из строя, да и получающиеся нефтепродукты, особенно мазут, будут низкого качества. Поэтому уже давно на нефтеперерабатывающих заводах поставлены специальные электрообессоливающие установки (ЭЛОУ), и нефть, прежде чем поступить на первичную переработку, обессоливается.

Принципиальная схема ЭЛОУ проста. Дело в том, что вода не растворяется в нефти — маленькие капельки воды, размерами от нескольких микрон до долей миллиметра, взвешены в нефти. Именно в этих капельках растворены соли. Капельки под действием силы тяжести осаждаются на дно трубопроводов или емкостей, в которых содержится нефть, образуя придонный слой воды, и эта вода может быть отобрана. Скорость осаждения капелек пропорциональна квадрату их линейного размера. Таким образом, если укрупнить капельки воды, то они будут достаточно быстро осаждаться, отделяться от нефти. Для этого нефть «промывают»: в нее добавляют большое количество слабо минерализованной — пресной воды с тем, чтобы маленькие капельки сильно минерализованной пластовой воды, сливаясь с капельками промывочной несоленой воды, образовывали капельки крупного размера, которые уже будут быстро осаждаться.

Думаю, вам ясно, что капелька, образовавшаяся от слияния пресной и соленой капелек, будет иметь меньшую соленость — удельную концентрацию солей в капельке — чем исходная соленая капелька. Для интенсификации процессов слияния капелек — коалесценции и осаждения воды в аппаратах ЭЛОУ нефть с водой пропускают через электрическое поле высокой напряженности. Нефть электрически нейтральна, а соленые капельки заряжаются, и это приводит к интенсификации процесса коалесценции, и осаждение капелек воды вместе с растворенной в них солью происходит быстрее.

На всех нефтеперерабатывающих заводах нашей страны установлены ЭЛОУ. Лет десять тому назад ситуация была следующая: на заводы поступала нефть с содержанием солей от нескольких сот до нескольких тысяч миллиграммов на литр, а после ЭЛОУ нефть содержала 20—40, а то и 60 миллиграммов солей на литр. Как видите, успехи были значительные, но все же и при таком, казалось бы, небольшом количестве солей их вредоносное действие еще очень велико. Народное хозяйство несло из-за этого многомиллионные потери. Поэтому было необходимо еще

в несколько раз снизить количество остаточных солей, то есть солей, остающихся в нефти после обработки на ЭЛОУ.

Таково было положение дел, когда лаборатории прикладной математики, которой я руковожу, было поручено заняться оптимизацией процесса на ЭЛОУ и обеспечить снижение количества остаточных солей до 5 миллиграммов на литр и меньше. Задача оптимизации ЭЛОУ на первый взгляд не казалась нам слишком трудной, и вот почему.

Обычная картина создания нового технологического процесса выглядит так: инженеры-технологи, будь то химики, нефтепереработчики, металлурги, пищевики выдумывают процесс, отрабатывают его на лабораторной установке или на малой промышленной (пилотной) установке; проектировщики создают промышленную конструкцию, и затем объект строят, часто несколько лет, и, наконец, запускают на заводе. Установка оснащена большим количеством контрольно-измерительных приборов и регуляторов, отработан технологический режим.

Выбранный режим найден эмпирически, без серьезной физико-химической теории или скрупулезной статистической обработки и обычно далек от оптимального: создание и проектирование сложного процесса — дело весьма нелегкое, и технологи не строят математическую модель процесса, годную для управления — слишком много параметров у процесса, и до реализации установки, как говорят, в металле трудно предвидеть детали ее поведения. Но когда установка запущена и функционирует нормально, возникает задача оптимизации процесса: нужно повысить выход целевого продукта, снизить непроизводительные затраты топлива, электроэнергии, дорогостоящих катализаторов и других веществ, используемых в производстве, сократить сроки выпуска продукции и т. д. Если выбранный технологический режим далек от оптимального, то еще есть возможность выбрать лучший режим, оптимизировать установку. Именно на это мы и рассчитывали, взявшись за оптимизацию процесса ЭЛОУ.

Для создания системы оптимального управления или хотя бы улучшения качественных характеристик процесса нужна его математическая модель. Впрочем, промежуточным этапом построения системы оптимального управления будет и определение некоторых закономерностей функционирования установки, выявление связей между параметрами, словом, изучение действующей установки.

Вследствие сложности процесса, как я уже сказал, построение математической модели, учитывающей всю гамму физико-химических процессов, обычно не удается, и возникает проблема построения модели по данным, полученным в условиях нормального функционирования интересующего нас объекта, годной для управления объектом в оптимальном режиме или в режиме, близком к оптимальному. Таким образом, модель должна хорошо описывать процесс, быть его двойником в системе управления.

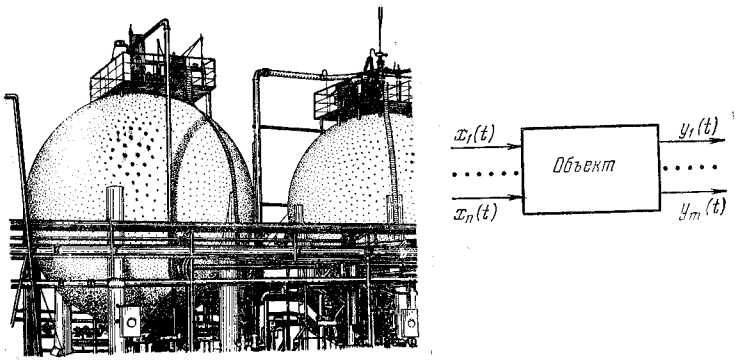


Рис. 8.

Но моделей можно построить много, и возникает проблема сопоставления модели и объекта для проверки того, годится ли выбранная модель на роль двойника — проблема идентификации, отождествления модели с объектом — оригиналом. Конечно, слово «отождествление» не нужно здесь понимать буквально, как при идентификации преступников — либо убийца тот же самый, либо другой. Никакая математическая модель не может абсолютно точно повторить объект во всем его многообразии. И поэтому речь идет о выборе среди математических моделей определенного класса той, которая наилучшим образом (в смысле принятого критерия) описывает поведение изучаемого объекта. Сам объект представляется в виде «черного ящика», то есть учитываются лишь значения некоторых его входов и выходов и не используется информация о состояниях самого объекта. Говоря другими словами, в модели не используется информация о процессах, происходящих внутри установки, и о значениях других

выходов и входов, которые исследователь полагает либо фиксированными, либо считает их вклад достаточно малым.

На рис. 8 вы видите реальную установку электрообессоливания (ЭЛОУ) и ее модель в виде «черного ящика» с теми же входами и выходами.

Схему рис. 8 можно упростить, если понимать под $x(t)$ и $y(t)$ вектор-функции соответственно с координатами $x_1(t)$, $x_2(t)$, ..., $x_n(t)$ и $y_1(t)$, $y_2(t)$, ..., $y_m(t)$. Тогда схема будет выглядеть совсем просто (рис. 9).

Наиболее широкое применение в механике, автоматике, радиотехнике при моделировании динамических систем нашли дифференциальные уравнения.

Рис. 9.

Первоначально задача идентификации решалась как задача определения коэффициентов в подобном уравнении по экспериментальным данным, собранным при реальной эксплуатации.

Однако при идентификации сложных систем нельзя всегда считать, что связь выхода и входа задана в виде линейного дифференциального уравнения или вообще какого-либо дифференциального уравнения. Выбор типа модели — это одна из основных и самых трудных задач моделирования. В действительности объект, схематизированный, как на рис. 8 или 9, осуществляет преобразование функции $x(t)$ в функцию $y(t)$, и модель так или иначе указывает это преобразование.

Но вернемся к проблеме электрообессоливания нефти. Чем меньше концентрация солей на входе ЭЛОУ, тем она меньше и на выходе — вот основной тезис нефтепереработчиков, и с позиций здравого смысла он выглядит вполне правдоподобно. Из этого тезиса следует вывод: для уменьшения в несколько раз концентрации солей на выходе ЭЛОУ надо снизить в несколько раз концентрацию солей на входе установки, то есть нужно еще на нефтепромыслах, вдали от производственных центров, создать мощные промышленные установки для обессоливания нефтей. Сделать это можно, но потребуются огромные капиталовложения. Однако, может быть, можно обойтись более простыми средствами — оптимизировать работу ЭЛОУ на заводах и тем обеспечить высококачественное обессоливание?

Если такое решение проблемы обессоливания возможно, то стоит над ним поработать, ибо оно обеспечит экономию огромных средств, которые в противном случае надо израсходовать на строительство стационарных промысловых установок.

Работу вести было трудно: мы почти всюду наталкивались на пессимизм нефтепереработчиков — они не верили в возможность существенного улучшения работы

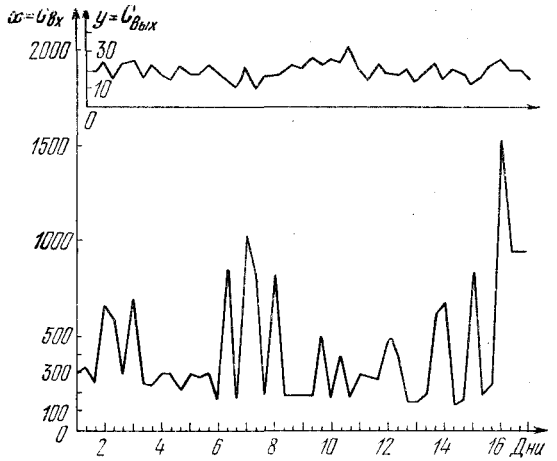


Рис. 10.

ЭЛОУ и считали необходимым снижение количества солей на промысле — уповали на тезис, отмеченный выше разрядкой.

Прежде чем давать рекомендации по оптимизации установки, надо знать, как же она функционирует.

Обозначим через x концентрацию $C_{вх}$ солей на входе установки и через y их концентрацию $C_{вых}$ на выходе. С интересующей нас стороны ЭЛОУ преобразует концентрацию солей x в концентрацию y в каждый момент времени, но функция, описывающая это преобразование, неизвестна.

Поэтому мы в первую очередь занялись сбором данных нормальной эксплуатации на полутора десятках установок разных нефтеперерабатывающих заводов. Эти данные были взяты из режимных листов — документов, заполняемых оператором, в которые вносятся значения измеряемых технологических параметров, в том числе и интересовавших нас концентраций солей на входе и выходе ЭЛОУ.

Вы видите на рис. 10 данные по одной из обследованных установок. Резко переменчивый, «истерический» характер нижней кривой — концентрации солей на входе — это следствие нерегулярности работы обезвоживающих промысловых установок, поступления в транспортный трубопровод нефтей не только с различных горизонтов или скважин, но и с разных месторождений и многих других причин. Обеспечить постоянство концентрации солей на входе или ее плавное изменение не представляется возможным.

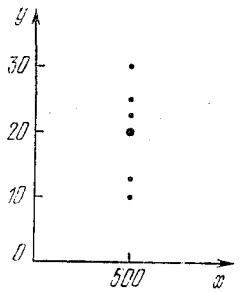


Рис. 11.

Сопоставим данные верхней и нижней кривых, рассматривая их значения в соответствующие моменты времени. Можно ли утверждать, поглядев на рис. 9, что между верхней и нижней кривыми есть функциональная зависимость, о которой нам говорили специалисты — нефтепереработчики: чем меньше концентрация солей на входе ЭЛОУ, тем меньше и на выходе? Здесь непосредственно, конечно, ничего не увидишь, и нужно, выбрав разумную математическую модель, подвергнуть имеющиеся данные обработке.

Для проверки правильности гипотезы нефтепереработчиков произведем следующее. Зафиксируем какую-то величину концентрации солей на входе ЭЛОУ, скажем 500 мг/л, и отберем все данные с такой входной концентрацией. Если бы гипотеза была верна точно, то и данные на выходе ЭЛОУ при 500 мг/л на входе были бы одинаковы. Но на самом деле они далеко не одинаковы, и на рис. 11 вы видите их различные значения.

Удивляться здесь нечему, и легко указать причины такого разнообразия: все подобные уклонения возникают вследствие большого числа помех, всегда присутствующих в любом процессе, а тем более таком «зашумленном» как ЭЛОУ, где есть много не учитываемых параметров и электрического, и гидродинамического, и технологического характера. Для избавления от таких помех, естественно, нужно усреднить, взяв, вместо имеющегося разнообразия данных, его математическое ожидание. Так как

математическое ожидание значений y берется при фиксации величины x=500 мг/л, то это условное математическое ожидание. Теперь все было бы в порядке, но для вычисления математического ожидания надо знать распределение вероятностей изучаемой случайной величины,

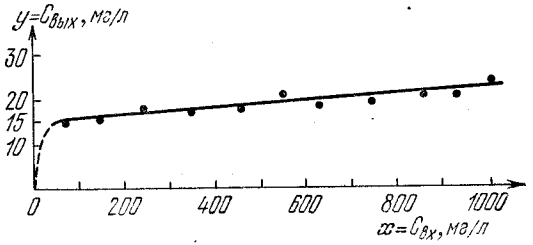


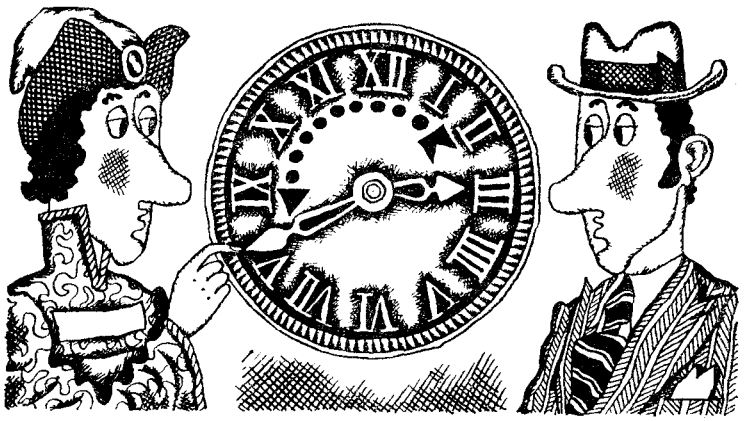
Рис. 12.

которое нам неизвестно: в нашем распоряжении лишь набор наблюденных значений — выборка. Как обычно в подобной ситуации, мы берем здесь среднее арифметическое и получаем условное эмпирическое среднее значение — оно соответствует жирной точке. Дальнейшая процедура, я надеюсь, вам понятна. Сначала для различных значений входных концентраций x строятся соответствующие точки y, и мы получаем на плоскости (x, y) облако данных, а затем при каждом значении x, то есть на каждой вертикали, усредняем имеющиеся данные и получаем набор жирных точек (рис. 12). Эти жирные точки и будут эмпирическим аналогом условного математического ожидания выхода по входу. Вот теперь и следует выяснить, имеется ли какая-то четкая связь между выходом y и входом x, или, говоря о нашей конкретной задаче, связь между концентрациями солей на входе и выходе.

В нашем случае эти данные оказались и весьма простыми, и в то же время малопонятными. Похоже, что точки лежат на прямой — можно провести прямую, к которой точки примыкают весьма близко. Ее, конечно, и следует взять в качестве математической модели связи выхода со входом для изучаемой установки. Некоторый разброс полученных нами эмпирических средних значений вполне объясним: данные получены в режиме нормальной эксплуатации установки, их количество ограничено, и проведенное усреднение не может полностью избавить нас от погрешностей. Но заметьте теперь, сколь маленьким является наклон прямой — она же идет почти параллельно горизонтальной оси, и, следовательно, установка оказывается весьма мало чувствительной к изменению концентраций солей на входе: изменение концентрации на входе с 1000 мг/л до 100 мг/л, то есть в 10 раз, дает едва заметные изменения выходной концентрации с 24 до 16 мг/л, то есть в 1,5 раза, причем величина в 1 мг/л лежит в пределах точности измерений. Аналогично выглядели данные по другим установкам. Как же трактовать такую зависимость? Если вдуматься, то вывод один: установки мало чувствительны к входной концентрации солей, они работают плохо и, даже при небольшой концентрации солей на входе, порядка 100 мг/л, не дают заметного уменьшения концентрации на выходе. Следовательно, тезис нефтепереработчиков — для уменьшения концентрации солей на выходе ЭЛОУ нужно заметно снизить концентрацию солей в поступающей на переработку нефти — оказался ошибочным.

Итак, с наших общих позиций было сделано следующее: построена математическая модель ЭЛОУ, отражающая связь концентрации солей на входе и выходе установки. При гипотезе о случайной структуре помех условное математическое ожидание выхода по входу дает линейное уравнение. Прямые линии хорошо, адекватно описывают имеющиеся экспериментальные данные. Но если присмотреться внимательно к полученным графикам, то заметно одно противоречие: на участок оси абсцисс от нуля до 100 мг/л продолжить прямую нельзя, ибо на установке нельзя получить концентрацию солей на выходе, скажем, 10 мг/л, когда на вход поступает нефть с концентрацией солей 5 мг/л,— появиться новым слоям здесь неоткуда. Поэтому на участке $0 \leqslant x \leqslant 100$ построенная модель в виде прямой линии не годится, нужно строить другую модель. Очевидно, функция должна проходить через начало координат: если нефть на входе ЭЛОУ совсем не содержит солей, то и на выходе солей не будет. Экспериментальных данных реальной эксплуатации на участке $0 \leqslant x \leqslant 100$ у нас не было, и пришлось строить модель, исходя из некоторых качественных физических соображений. На этом участке мы ее выбрали в виде нарастающей экспоненты, на рис. 12 она отмечена пунктиром.

Я, по-видимому, переутомил вас подробным разбором этого конкретного примера, но сделал это не из желания рассказать знакомую работу,— здесь мы с вами не только проследили последовательность действий при построении математической модели для идентификации сложного объекта, но и увидели, зачем такая модель может быть нужна, какую пользу можно извлечь, тщательно продумывая результаты идентификации объекта. Дальнейшее продумывание полученных результатов показало, что нельзя осуществить никакую оптимизацию за счет изменения характеристик или параметров действующей ЭЛОУ, имеющихся, так сказать, в наших руках. Следовательно, оказалось необходимым тщательно разобраться в физике и даже в физической химии процесса электрообессоливания. Работа оказалась трудной, но в результате мы не только поняли физико-химические закономерности весьма сложного процесса. Оказалось, что для заметного повышения эффективности процессов коалесценции и дробления капель необходимо значительно увеличить время пребывания эмульсии в электрическом поле. Для этого было сконструировано и запущено в действие специальное устройство — электрокоалесцентор, обеспечивающее необходимое время пребывания эмульсии в электрическом поле. Кроме того, более глубокое понимание физико-химических закономерностей процесса привело к необходимости изменения дозировки объемов промывочной воды, объемов и места подачи деэмульгатора, уточнения других важных технологических величин, дало возможность наладить процесс и в результате снизить остаточное содержание солей в 3—4 раза, что привело к очень значительной экономии материальных средств, дефицитных реагентов и пресной воды.



Регрессия

Вернемся к облаку данных вход — выход. Аналогичное облако получится, если координаты точек будут представлять собой длину x и диаметр y берез в роще или соответственно длину x и ширину y головы преступников при бертильонаже. Таким образом, сейчас уже можно отвлечься от реального содержания величин x и y и поставить общую задачу.

Имеются две случайные величины ξ и η, между которыми есть некоторая зависимость: похоже, что чем большее значение x принимает ξ, тем большее значение y принимает η, или значение y становится меньше при возрастании x, или зависимость между x и y выражается как-то иначе, в виде квадратичной функции, или колебательной, или еще как-то. Неопределенность моих высказываний объясняется отсутствием четкой, строго определенной, детерминированной зависимости — зависимость эта статистическая, и если я говорю о наличии зависимости, то она видна лишь в среднем: если нарисовать облако данных, соответствующее реализациям случайных величин (ξ, η), то есть наблюденным парам значений — точкам (x_i, y_i), $i=1, \ldots, n$, то они располагаются вдоль какой-то кривой. На рис. 13 представлена подобная ситуация, когда, в отличие от рис. 12, точки концентрируются вдоль кривой, имеющей выраженный максимум и минимум.

Теоретически довольно просто найти эту кривую, если пара (ξ, η) задана своим совместным распределением вероятностей: тогда-то и следует взять в качестве кривой условное математическое ожидание случайной величины η при условии, что случайная величина ξ приняла значение x:

$$y = \mathsf{M}(\eta \mid \xi = x) = \varphi(x).$$

Эта функция и будет искомой нами зависимостью «в среднем» между η и ξ. Уравнение $y=\varphi(x)$ называется уравнением регрессии или, более точно, уравнением регрессии η на ξ, ибо можно рассматривать и уравнение

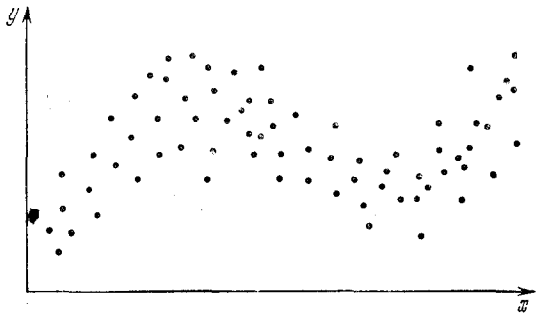


Рис. 13.

$x=\mathsf{M}(\xi \mid \eta=y)=\psi(y)$ — уравнение регрессии ξ на η, причем графики функций $y=\varphi(x)$ и $x=\psi(y)$, вообще говоря, не совпадают.

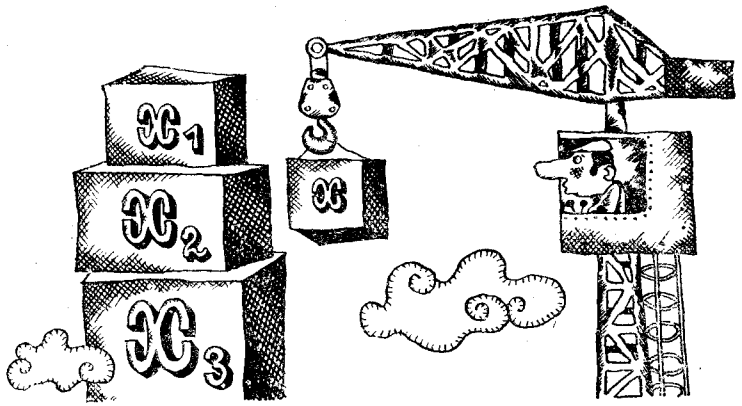
Слово регрессия в статистику ввел Френсис Гальтон, один из создателей математической статистики. Сопоставляя рост детей и их родителей, он обнаружил, что соответствие между ростом отцов и детей слабо выражено, оно оказалось меньшим, чем он ожидал. Однако Гальтон не унывал — он объяснил это явление наследственностью не только от родителей, но и от более отдаленных предков: по его предположениям, то есть по его математической модели, рост определяется наполовину родителями, на четверть — дедом и бабкой, на одну восьмую — прадедом и прабабкой и т. д. Я не знаю, прав ли здесь Гальтон, но он обратил внимание на движение назад по генеалогическому дереву и назвал это явление регрессией, заимствовав понятие движения назад, противоположное прогрессу — движению вперед. Впоследствии слово «регрессия» заняло в статистике заметное место, хотя, как это часто бывает в любом языке, в том числе и в языке науки, в него теперь вкладывают другой смысл — оно означает статистическую связь между случайными величинами.

На практике мы почти никогда не знаем точный вид распределения, которому подчиняется величина, а тем более вид совместного распределения двух или большего количества случайных величин, и поэтому уравнение регрессии $y=\varphi(x)$ нам тоже неизвестно. В нашем распоряжении лишь некоторый набор наблюдений — облако данных, ну и возможность строить модели уравнения регрессии, а затем проверять их, опираясь на эти данные. Как вы видели, при построении модели облака данных для зависимости вход — выход в процессе электросбессоливания в качестве математической модели было совершенно естественно взять линейную функцию. Ее можно записать в виде

$$y=\beta_0+\beta_1 x+\varepsilon,$$

где β_0 и β_1 — коэффициенты, которые нужно определить по экспериментальным данным, а ε — ошибка, которую мы полагаем случайной, имеющей нулевое математическое ожидание и такой, что ее значения в различных точках (x_i, y_i) будут независимыми.

Если же облако выглядит иначе, например так, как на рис. 13, где прямая линия ему плохо соответствует, то возникает проблема подбора функции — математической модели для неизвестного нам уравнения регрессии.



Из кубиков дети складывают различные фантастические замки.

Менее романтические взрослые из кирпичей строят дома и заводы, санатории и скотные дворы. Жилые дома не должны быть одинаковыми и по внешнему оформлению, и по «начинке»: разным семьям нужны разные квартиры, противно ходить среди домов — близнецов, различающихся лишь окраской балконов. Следует учитывать и климатические особенности — в Архангельске и Ташкенте нужно строить по-разному.

Конечно, дом в любом стиле от готики до конструктивизма можно построить из кирпичей, но строить из них невыгодно — дешевле и быстрее складывать здания из панелей и целых блоков, то есть воспользоваться набором стандартных деталей — их почему-то называют типовыми. И естественно возникает задача: как составить сам набор типовых деталей, из которых можно быстро, хорошо и дешево складывать разнообразные строения. Не будем обсуждать варианты решения этой важной проблемы — я упомянул о ней, имея лишь в виду ассоциации.

Основной объект математического анализа — функция, то есть зависимость одной переменной от другой, или от других, когда независимых переменных несколько. Различных мыслимых функций столь много, что попадаешь в безнадежное положение, когда пытаешься представить

себе все их разнообразие... К счастью, инженеру, биологу или экономисту в этом нет нужды: ему не нужны абсолютно точные сведения о поведении функции при изменении независимых переменных. Действительно, строителю достаточно знать размеры и графики профилей с точностью до миллиметра, инженер-радист обычно использует график вольт-амперной характеристики лампы с относительной точностью в несколько процентов, врача устраивает знание температурной кривой пациента с точностью до одной десятой градуса. Даже траекторию космического корабля нужно знать хотя и с высокой, но ограниченной точностью.

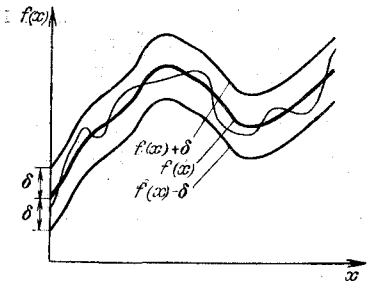


Рис. 14.

Точность, с которой нужно знать функцию в каждой точке,— это некоторое число δ, конечно, положительное. Оно может быть равно трем десятым или одной сотой или двадцати — точность зависит от условий задачи, наших возможностей или желаний.

Представим себе наглядно всю картину: на чертеже (рис. 14) вокруг графика функции надо нарисовать полоску, ширина которой в направлении вертикальной оси будет 2δ. Для этого сдвинем, как целое, график функции вверх и вниз на δ, и тогда часть плоскости между верхним и нижним графиками будет нужной δ-полоской.

Любая кривая, целиком лежащая внутри δ-полоски, например нарисованная тонко, с точностью до δ неотличима от исходной, нарисованной жирно.

Теперь возникает вопрос: если функцию не надо знать абсолютно точно, если можно пренебречь мелкими деталями, то нельзя ли заменить произвольную функцию какой-либо близкой к ней, но более простой, легче поддающейся изучению?

Вы, конечно, знаете, что ответ на этот вопрос положительный; впрочем, иначе зачем бы я строил всю эту подготовительную паутину.

Но мне хочется разъяснить нечто более глубокое. Оказывается, функции, все более точно приближающие инте-

ресующую нас данную (хотя и любую) функцию, можно складывать из простых функций — кирпичиков. Для определенности я буду вести рассказ о непрерывных функциях. Их графики не имеют разрывов — непрерывные функции удобно представлять себе похожими на ниточку. Разнообразие непрерывных функций огромно. Например, существуют непрерывные функции, не имеющие ни в одной точке производной, то есть не обладающие ни в одной точке касательной к кривой. Представить себе наглядно подобную функцию трудно. Но с позиций практика такие функции не представляют интереса — их нельзя реализовать ни в какой физической системе.

Среди множества всех непрерывных функций есть хорошо вам известные — многочлены. Я приношу извинения за банальность, но напомню формальную запись многочлена n-й степени:

$$P_n(x) = a_0 + a_1 x + \ldots + a_n x^n.$$

У него имеется $n+1$ коэффициентов $a_0, a_1, \ldots, a_n$ — это произвольные вещественные числа. Нетрудно приближенно построить график конкретного многочлена, у которого заданы численные значения коэффициентов вроде

$$P_9(x) = 12 - 4x + 18x^5 - 0{,}01x^9.$$

Для этого, придавая различные числовые значения независимой переменной x в интересующем нас интервале, нужно их подставить в формулу и вычислить соответствующую алгебраическую сумму.

В общем, едва ли многочлены кажутся вам такими уж сложными функциями. Как вы представляете себе их возможное разнообразие? Попробуйте, прежде чем читать дальнейшее, поразмышлять об этом.

Среди многочленов простейшие — степенные функции $x, x^2, \ldots, x^n, \ldots$. Добавим еще сюда и 1 — степенную функцию нулевой степени.

Взяв x^3 пять раз, затем умножив x^2 на (-2) и сложив с предыдущим, прибавив далее x, умноженный на (-3), и прибавив, наконец, взятую пять раз 1, мы получим многочлен

$$P(x) = 5 - 3x - 2x^2 + 5x^3.$$

Таким образом, любой многочлен можно сложить из степенных функций — кирпичиков. Впрочем, если вы считаете, что все степенные функции — разные, то их следует назвать не кирпичиками, а типовыми деталями, и построение многочленов из этих деталей с архитектурной точки зрения выглядит вполне современно.

Вот теперь мы подобрались к самому интересному. Возьмем произвольную непрерывную функцию на выбранном отрезке $0 \leqslant x \leqslant 1$ и устроим вокруг нее δ-полоску. Как мы уже установили, любая другая непрерывная функция, график которой расположен целиком внутри той же δ-полоски, неотличима с точностью до δ от исходной.

Оказывается, среди функций, графики которых целиком лежат внутри δ-полоски, найдется и многочлен.

Я хочу еще раз подчеркнуть и парадоксальность, и фундаментальность этого факта, сказав то же самое другими словами: как бы сложно (угловато, с резкими колебаниями, подъемами или спадами) ни была устроена непрерывная функция, и какое маленькое δ мы бы ни выбрали, найдется многочлен, не отличимый с точностью до δ от этой конкретной непрерывной функции.

Таким образом, мой вопрос о том, как вы себе представляете разнообразие многочленов, довольно ехидный. Но не корите себя, если вам показалось множество многочленов проще, чем оно есть на самом деле. Утверждение, выделенное разрядкой,— одна из самых фундаментальных теорем математического анализа; оно носит название теоремы Вейерштрасса об аппроксимации (приближении) непрерывных функций многочленами.

Карл Вейрштрасс (1815—1897) — один из наиболее крупных математиков девятнадцатого века. Он получил фундаментальные результаты почти во всех разделах математического анализа. Более того, среди плеяды великих математиков девятнадцатого века, осуществивших перестройку математики на новой более строгой и совершенной основе, имя Вейерштрасса занимает одно из первых мест.

Но вернемся к теореме Вейерштрасса. Ее можно интерпретировать еще и так. Возьмем конкретную, но совершенно произвольную непрерывную функцию $f(x)$ и какую-либо последовательность убывающих и стремящихся к нулю чисел, например $\delta_1 = 10^{-1}$, $\delta_2 = 10^{-2}, \ldots, \delta_n = 10^{-n}, \ldots$ В соответствии с обсуждаемой теоремой можно для каждой из этих величин δ_n подобрать многочлен, который с точностью до δ_n будет неотличим от приближае-

мой функции $f(x)$. Если обозначить многочлены соответственно через $P_1(x)$, $P_2(x), \ldots, P_n(x), \ldots$, то получается последовательность многочленов, все более тесно примыкающих к функции $f(x)$, и так как последовательность чисел δ_n стремится к нулю с возрастанием номера, в пределе последовательность многочленов и дает нашу функцию $f(x)$.

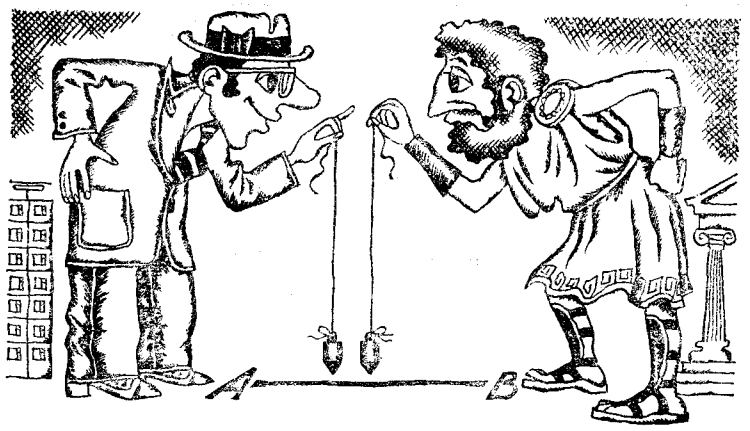
Таким образом, последовательность приближающих многочленов потенциально содержит все свойства исходной функции $f(x)$.

Как видите, именно теорема Вейерштрасса указывает на принципиальную возможность практику избавиться от давящего многообразия всех непрерывных функций и, когда это целесообразно, оперировать лишь с многочленами.

Конечно, чем выше будет требуемая точность приближения функции многочленом (то есть чем меньше будет δ), тем, вообще говоря, более высокой окажется степень приближающего многочлена. Но все же многочлены значительно легче поддаются изучению, чем произвольные непрерывные функции.

В строительстве набор типовых деталей должен меняться в зависимости от конкретных условий: едва ли целесообразно крупные заводские корпуса складывать из тех же панелей, что и дома с малогабаритными квартирами.

Так же и в математике: многочлены — далеко не единственные «типовые детали», из которых можно сложить функции, приближающие произвольную непрерывную функцию с любой заданной точностью.



Обратимся к геометрии

Геометрические представления обычного трехмерного пространства очень наглядны, и при построении более общих пространств широко пользуются аналогиями с трехмерным пространством: многие факты евклидовой геометрии остаются верными и для многомерных пространств, а те, что требуют уточнений, отправляются от привычных геометрических представлений. Поэтому сохраняется в основном и терминология.

Я сейчас остановлюсь на некоторых понятиях теории многомерных пространств. Часть из них, конечно, требует точных формулировок и доказательств. Но для пояснения геометрического смысла некоторых фактов, полезных для дальнейшего, можно обойтись аналогиями и наводящими рассуждениями, и поэтому я буду говорить не очень точно, надеясь на вашу геометрическую интуицию.

Если элементы теории многомерных пространств, включая бесконечномерные, вам известны, то не теряйте зря времени, которого всегда не хватает, и пропустите несколько страниц.

Вы, конечно, помните, что вектор — это стрелка, иначе говоря, отрезок прямой, у которого есть длина и направление. Угол между векторами удобно задавать через скалярное произведение. Напомню: если x и y — векторы, то их скалярное произведение (x, y) равно произведению

их длин на косинус угла между ними. При построении теории многомерных векторных пространств удобно в качестве исходного понятия взять именно скалярное произведение, причем задать его с помощью аксиом. Я не буду их перечислять — они сейчас не понадобятся. Но важно отметить, что квадрат длины вектора равен скалярному произведению вектора с самим собой:

$$(x, x) = \|x\|^2,$$

где $\|x\|$ — это обозначение длины или, как говорят, нормы вектора. Угол α между векторами x и y задается тоже через скалярное произведение:

$$\cos\alpha = \frac{(x, y)}{\|x\| \cdot \|y\|}.$$

Если угол между векторами x и y прямой, то их скалярное произведение равно нулю. Такие векторы называются ортогональными, а в элементарной геометрии — перпендикулярными.

Множество векторов, которые можно складывать (по обычному правилу параллелограмма) и умножать на числа, образует линейное векторное пространство. Оно может быть не только двумерным — плоскостью или трехмерным, как наше обычное пространство, но может иметь любое число измерений. Число измерений — размерность пространства — определяется наибольшим количеством взаимно ортогональных векторов, которые можно поместить в это пространство. Такие векторы — их совокупность называют ортогональным базисом — естественно принять за оси координат, и тогда каждый вектор разлагается по осям координат. Если $x_1, x_2, \ldots$ — это величины проекций вектора x на координатные векторы $e_1, e_2, \ldots$, то имеет место обобщение теоремы Пифагора: длина вектора $\|x\|$ выражается формулой

$$\|x\|^2 = \sum_k x_k^2. \qquad (*)$$

Если пространство конечномерно, то есть наибольшее число взаимно ортогональных векторов конечно, то формула (*) достаточно очевидна. Но можно рассматривать и бесконечномерные линейные векторные пространства. В таком пространстве имеется бесконечное количество взаимно ортогональных векторов, и тогда в формуле (*) нужно предположить, что ряд сходится. В этой ситуации

пространство называется гильбертовым в честь знаменитого немецкого математика Давида Гильберта (1862—1943), который в работах в 1904—1910 годов впервые использовал геометрические идеи бесконечномерного пространства в теории интегральных уравнений.

При аксиоматическом построении линейного векторного пространства и, в частности, гильбертова пространства от элементов пространства — векторов не требуется ничего, кроме возможности их складывать и умножать на числа и выполнения некоторых аксиом, среди которых следует подчеркнуть аксиомы скалярного произведения.

В такой интерпретации в качестве векторного пространства можно рассматривать весьма разнообразные множества элементов. В частности, множество функций, заданных на отрезке, также образует линейное векторное пространство, если скалярное произведение задать в виде интеграла от их произведения:

$$(x, y) = \int_a^b x(t)\, y(t)\, dt. \qquad (**)$$

В это пространство входят, конечно, не все мыслимые функции, а лишь те, для которых существует интеграл от квадрата функции — квадрат длины вектора:

$$(x, x) = \|x\|^2 = \int_a^b x^2(t)\, dt.$$

Пространство всех таких функций также называется гильбертовым и обозначается $L_2(a, b)$.

Замечу теперь, что если два вектора x_1 и x_2 в векторном пространстве не параллельны, то множество всех их линейных комбинаций $a_1x_1+a_2x_2$, где a_1 и a_2 — произвольные числа, заполняет плоскость. Соответственно линейные комбинации n векторов x_i вида $a_1x_1+a_2x_2+\ldots+a_nx_n$, когда коэффициенты a_i пробегают все множество вещественных чисел, заполняют, вообще говоря, n-мерное пространство. Говорят, что оно натянуто на векторы x_1, $x_2, \ldots, x_n$.

Если размерность исходного пространства строго больше n, то построенное n-мерное пространство называется подпространством исходного пространства. В гильбертовом пространстве — бесконечномерном — любое простран-

ство, натянутое на n его векторов, будет подпространством. Однако в гильбертовом пространстве есть и бесконечномерные подпространства, например подпространство, натянутое на все координатные векторы, кроме первых трех, или натянутое на все координатные векторы с нечетными номерами.

Обратимся к одной простой задаче элементарной геометрии. В обычном трехмерном пространстве — назовем его R — рассмотрим плоскость Q, проходящую через начало координат, и вектор y, не лежащий в этой плоскости. Как найти в плоскости Q вектор, самый близкий к вектору y? Вы, конечно, знаете ответ: нужно из конца вектора y опустить перпендикуляр на плоскость Q, и получившийся при этом вектор $\hat{y}$ — проекция вектора y на плоскость Q — будет самым близким к y из всех векторов плоскости. Иначе говоря, наилучшим приближением вектора y с помощью векторов из плоскости Q будет вектор $\hat{y}$ — проекция вектора y на плоскость (рис. 15).

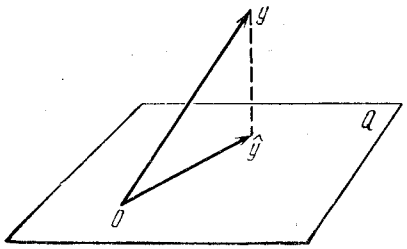


Рис. 15.

Та же задача о наилучшем приближении возникает и в теории многомерных пространств: если y — вектор пространства H (любой размерности) и Q — его подпространство, не содержащее вектора y, то наилучшим приближением вектора y с помощью векторов из подпространства Q будет вектор $\hat{y}$ — проекция вектора y на подпространство Q.

Можно указать и ошибку наилучшего приближения: она, очевидно, равна длине $\|y-\hat{y}\|$ перпендикуляра, опущенного из конца вектора y на подпространство Q.

Если подпространство Q натянуто на векторы x_1, x_2,..., x_m, то опустить перпендикуляр из конца вектора y на Q — это значит найти вектор z, ортогональный каждому из векторов x_1, x_2,..., x_m. Такая задача легко сводится к решению системы линейных алгебраических уравнений.

В нашей геометрической интерпретации все выглядит очень просто. Но вспомним теперь о гильбертовом пространстве функций. Здесь векторы — это функции на отрезке, подпространство, натянутое на n векторов,— все возможные линейные комбинации этих функций, и обсужденная задача означает построение наилучшего приближения к некоторой функции с помощью указанных линейных комбинаций. В аналитических терминах задача о наилучшем приближении заданной функции с помощью линейных комбинаций других функций не выглядит такой простой, и геометрический подход указывает один из возможных способов ее решения, причем дает очень прозрачную картину всех необходимых при этом операций.

Обсужденная задача о наилучшем приближении — это другая, чем в теореме Вейерштрасса, постановка проблемы приближения функций линейными комбинациями каких-либо более простых, другая в том смысле, что в разных постановках используются различные понятия расстояния между функциями. В теореме Вейерштрасса за расстояние между функциями принимается наибольшее расстояние между их графиками в направлении вертикальной оси; оно берется по абсолютной величине. Здесь же расстоянием между функциями-векторами служит норма их разности

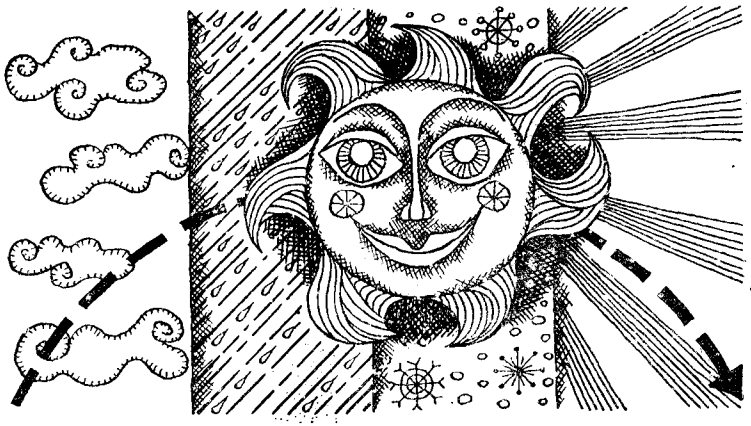
$$\|x-y\| = \sqrt{\int_a^b (x(t)-y(t))^2\,dt},$$

то есть корень квадратный из величины площади между горизонтальной прямой и графиком квадрата разности функций.

Когда на проблему представления функций в виде сумм некоторых «типовых функций» удалось посмотреть с более общих позиций, стало ясно: есть много систем функций, из которых можно сложить, как из кирпичиков, любую непрерывную функцию.

В качестве исходных функций, посредством которых приближают заданную, удобно взять последовательность взаимно ортогональных функций — ортогональный базис рассматриваемого гильбертова пространства функций.

Сказанное означает воможность представления каждой функции из пространства в виде линейной комбинации функций базиса. Таким образом, базисные функции — это как раз те типовые детали, из которых и складывается все разнообразие рассматриваемого функционального пространства.



Солнце всходит и заходит...

День сменяет ночь; за теплым летом следует осенняя промозглая слякоть и холоднющая зима, но, к счастью, наступает снова весна и за ней лето; сердце сокращается, скажем, семьдесят шесть раз в минуту.

Все это повторяющиеся, периодические процессы. Их периоды — соответственно сутки, год и $^1/_{76}$ минуты. Подобные периодические процессы встречаются повсеместно. Обнаружено, например, более сотни физиологических систем, функционирующих по законам суточной периодичности.

Одно из самых старых технических устройств, реализующих периодический процесс,— это маятник. Отведем маятник в сторону и отпустим — маятник будет качаться туда — сюда. Пренебрежем сейчас силой трения, замедляющей его движение, и проследим изменение во времени величины α — угла отклонения маятника от положения равновесия — от вертикальной прямой. Как вы наверное знаете из курса физики, величина α оказывается периодической функцией времени, и ее период — это период колебаний маятника, то есть интервал времени, за который маятник впервые вновь возвращается, скажем, в крайнее левое положение. Можно записать движение маятника формулой

$$\alpha(t) = A\cos\left(\frac{2\pi}{T}t + \varphi\right).$$

Здесь A — амплитуда колебаний, t — время, T — период колебаний маятника, φ — начальная фаза. Обычно вводят круговую частоту $\omega = 2\pi/T$, и тогда кривая на рис. 16 будет выражаться формулой $\alpha(t) = A\cos(\omega t + \varphi)$.

Периодические процессы весьма разнообразны. Приведу несколько примеров из разных областей.

Поговорка гласит: «На то и щука, чтоб карась не дремал». Когда в озере водятся щуки и нехищные рыбы,

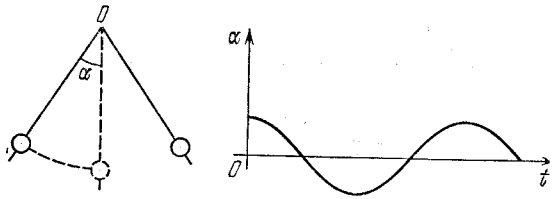


Рис. 16.

скажем караси, и карасей много, то щуки живут вольготно, интенсивно поедая карасей и быстро размножаясь. В результате количество карасей резко убывает, а щук — возрастает, и щукам постепенно становится нечего есть. Без пищи не проживешь, интенсивность размножения щук падает, и их поголовье убывает. Тогда для карасей наступает раздолье, и их стадо возрастает. Снова у щук появляется пища — и все начинается сначала...

Аналогичные периодические процессы наблюдаются в экономике. В условиях свободного предпринимательства происходят циклические колебания цен на продукты сельского хозяйства. Высокая цена на свиней служит для крестьян стимулом к интенсификации свиноводства. Вследствие этого, по данным одного немецкого экономиста, в двадцатых годах нашего века примерно через полтора года количество свиней на рынке заметно возрастало, и цена на них падала. С этого момента начинался обратный процесс — крестьяне сокращали воспроизводство свиней до момента, когда, уже вследствие недостатка свиней, цена на них вновь возрастала. Если никакие другие причины не нарушают течения такого процесса, то цена на свиней претерпевает колебания, похожие на синусоидальные с периодом примерно в три года.

Ряд периодических процессов тесно связан с сердечными сокращениями. Многие знают электрокардиограм-

му — запись биотоков, снимаемых с области, близкой к сердцу. На рис. 17 представлена электрокардиограмма, где явно видна периодичность всплесков — импульсов тока.

Для приближения периодических функций использовать в качестве кирпичиков многочлены не целесообразно — придется брать слишком высокие степени, и здесь в качестве естественных кирпичиков появляются гармо-

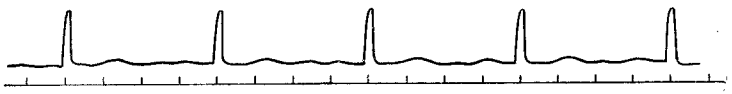


Рис. 17.

нические колебания кратных частот. Именно, если период функции равен T и соответственно круговая частота есть $\omega=2\pi/T$, то кирпичиками служат синусоидальные функции кратных частот: $1, \sin \omega t, \cos \omega t, \sin 2\omega t, \cos 2\omega t, \ldots, \sin n\omega t, \cos n\omega t$. Вместо многочленов теперь выступают тригонометрические многочлены вида

$$s(t)=5-2\sin t+0{,}3\cos 2t-0{,}1\cos 4t.$$

Как вы понимаете, для примера я написал какие попало численные значения коэффициентов, взяв частоту $\omega=1$.

Полезно записать тригонометрический многочлен общего вида

$$s(t)=a_0+a_1\cos\omega t+b_1\sin\omega t+\ldots+a_n\cos n\omega t+b_n\sin n\omega t,$$

где $a_0, a_1, b_1, \ldots, a_n, b_n$ — числовые коэффициенты.

Оказывается, любую непрерывную периодическую функцию можно с любой степенью точности приблизить тригонометрическим многочленом.

Эта теорема, столь же фундаментальная, как и предыдущая, также принадлежит Карлу Вейерштрассу. Впрочем, еще Жозеф Фурье в первой четверти девятнадцатого века в связи с исследованиями по теории теплопроводности весьма эффективно, да и эффектно использовал представление функций в виде сумм синусоидальных колебаний кратных частот. Поэтому ряды, представляющие функции в виде сумм гармонических колебаний, носят название рядов Фурье.

Если воспользоваться геометрическим подходом, о котором говорилось в предыдущем разделе, то здесь рассматривается та же задача о представлении функции с помощью линейных комбинаций базисных функций, и в качестве последних выбраны тригонометрические функции кратных частот. Остается лишь убедиться в их ортогональности. Для этого нужно вычислить несколько простых интегралов.

Сами же коэффициенты Фурье — скалярные произведения исходной функции с базисными тригонометрическими функциями — также просто выражаются с помощью интегралов. Все эти формулы есть в любом справочнике по математическому анализу.

Замечу кстати, что степенные функции x^m, о которых шла речь в разделе «Кирпичики», не являются попарно ортогональными. Однако можно среди многочленов выбрать и попарно ортогональные. Впервые ортогональные многочлены ввел в 1785 году один из крупнейших французских математиков конца XVIII — начала XIX века Адриен Мари Лежандр. Приведу первые пять многочленов Лежандра:

$$P_0(x)=1,\quad P_1(x)=x,\quad P_2(x)=\frac{1}{2}(3x^2-1),$$

$$P_3(x)=\frac{1}{2}(5x^2-3x),\quad P_4(x)=\frac{1}{8}(35x^4-30x^2+3).$$

Нетрудно проверить, что они ортогональны на отрезке $[-1, +1]$, то есть при $n\neq m$

$$\int_{-1}^{+1} P_n(x)\,P_m(x)\,dx=0.$$

Если речь идет об абсолютно точном представлении непрерывной функции в виде алгебраической суммы базисных, то ряды будут содержать, вообще говоря, бесконечное количество членов. Такое представление, конечно услаждает душу чистого математика и часто выглядит на самом деле очень красиво. Прикладнику же нужно исхитриться представить сложную функцию с необходимой ему точностью суммой небольшего числа простых функций.

Если же речь идет о представлении на некотором отрезке (a, b) непрерывной функции с точностью до δ, то такую задачу можно решить, взяв, например, в качестве приближения первые n членов ее ряда Фурье. Но количе-

ство функций в таком представлении может быть и весьма значительным: оно возрастает при уменьшении δ. Поэтому удачный выбор базисных функций, обеспечивающий в вашей задаче удовлетворительную точность приближения при малом числе базисных функций,— это очень важный элемент исследования. .

Вы, конечно, сейчас ждете от меня рецептов или хотя бы рекомендаций для выбора базиса, обеспечивающего и нужную точность, и малое количество базисных функций в сумме, и их простоту. К сожалению, как это обычно в прикладной математике, общих рекомендаций дать нельзя — удачный выбор базисных функций существенно зависит от задачи, имеющихся сведений об изучаемом объекте и наличии экспериментального материала. Скажем, если интересующая вас функция — это отклик линейной системы с постоянными параметрами (пассивный четырехполюсник, как говорят радисты), то естественно базис искать среди гармонических колебаний и экспонент, если же изучается отклик системы с изменяющимися параметрами (например, отклик электрического контура с изменяющейся емкостью), то базисными будут специальные функции, существенно зависящие от закона изменения параметров. Если вы с ними не знакомы, они вам покажутся достаточно сложными.

Вернемся к облаку данных на рис. 13 раздела «Регрессия». Уравнение регрессии $y=\varphi(x)$ нам неизвестно, и из экспериментального материала, независимо от обширности данных, нельзя его вывести. Напомню: сначала нужно выдвинуть гипотезу о виде зависимости или, говоря другими словами, выдумать математическую модель и лишь затем, пользуясь экспериментальными данными, проверять ее адекватность.

Поэтому исследователь, поглядев на облако из точек, покопавшись в своем арсенале ассоциаций и вспомнив различные методы построения математических моделей, выбирает вид модели. Конечно, нужно начинать с чего-нибудь наиболее простого: с линейной модели (то есть с прямой линии), синусоиды, параболы или какой-либо другой простой функции.

Где-то я прочитал такое высказывание: «Одной из основных задач теоретического исследования в любой области является установление такой точки зрения, с которой объект исследования проявляется с наибольшей простотой». Это, конечно, верно, если ясна разница между простым и сложным... Но простота представляется довольно условной категорией, она в значительной мере зависит от привычки, опыта, осведомленности. Мне рассказывали, как знаменитому человеку на семидесятилетний юбилей преподнесли торт и зажгли на нем семьдесят свечей. Он

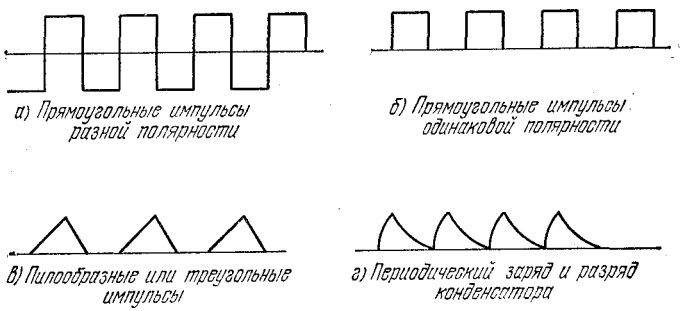


Рис. 18.

предложил своей трехлетней внучке потушить свечи, и девочка спросила: «А где выключатель?»

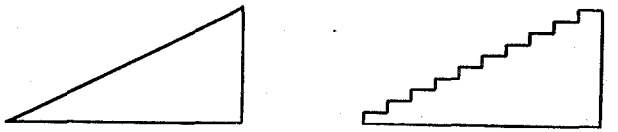
Синусы и косинусы у восьмиклассника вызывают недоумение и душевный трепет, особенно если их ввести формально, как отношение катетов к гипотенузе, и не объяснить их связь с колебаниями. В то же время электрику привычны синусоидальные периодические колебания с частотой 50 герц — это же частота промышленного переменного тока, который вы используете у себя дома. Телевизионщик имеет дело с импульсными периодическими процессами (см., например, графики на рис. 18).

Поэтому, если в качестве базисных функций в задаче, даже далекой от телевидения, будут выбраны импульсы с разной частотой повторения и разной амплитудой (высотой столбиков), то у специалиста-телевизионщика такое представление не вызовет ни недоумения, ни ощущения сложности — оно естественно для определенного класса процессов, с которыми телевидение постоянно имеет дело.

Но не следует использовать базис из прямоугольных импульсов для представления треугольного сигнала, который вы видите на рис. 18, *в*, — такое представление даже при незначительной точности потребует модели из большого числа базисных функций. На рис. 19 видно, о чем идет речь: треугольный импульс (рис. 19, *а*) при-

близительно заменен на сумму из десяти импульсов одинаковой продолжительности, но разной амплитуды (рис. 19, *б*), и получившаяся «лестница» все еще с малой точностью воспроизводит треугольный импульс.

Таким образом, выбор хорошего базиса требует и опыта, и ясного понимания физического существа решаемой



а) *б)*

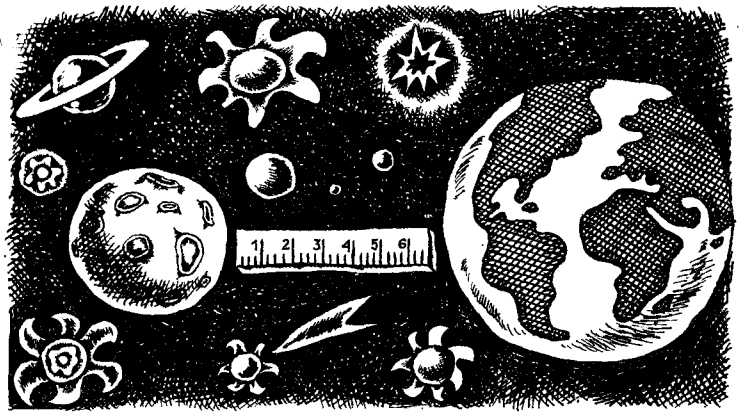
Рис. 19.

задачи. Впрочем, эти составляющие полезны на всех этапах исследования...

Но законы природы или фактические зависимости между переменными в разумной модели в физике, химии, биологии, экономике, социологии, как мне думается, не могут быть слишком сложными. Если реальные данные приводят к модели, удовлетворительно описываемой лишь многочленом сотого порядка, то, с позиций здравого смысла, либо нужно пересмотреть постановку задачи — она может быть плохо поставлена, что и влечет за собой неоправданные осложнения,— либо исследователь плохо выбрал систему базисных функций, и нужно строить совсем другую математическую модель изучаемого объекта или явления, складывать ее из других типовых функций, отвечающих и объекту, и поставленной задаче.

Моя вера в простоту и ясность законов природы и реальных зависимостей даже в сложных системах основана на опыте. И если вы строили математические модели реальных процессов или объектов, то согласитесь со мной, если же это вам придется делать в будущем, то скептическое отношение к витиеватым, малопонятным, громоздким, плохо обозримым построениям примите как руководство к действию, и надеюсь, такой скептицизм приведет вас к успеху.

Словом, прежде, чем пускаться в расчеты или эксперименты, нужно серьезно подумать, как же выбрать математическую модель,— от этого может существенно зависеть успех работы.



Самая близкая

Но вот, наконец, выбран вид математической модели, и остаются лишь неопределенными коэффициенты.

Например, для облака экспериментальных данных на рис. 13, выражающих неизвестную нам связь переменных x и y, кажется целесообразным попробовать модель в виде многочлена третьей степени:

$$y=\alpha_0+\alpha_1 x+\alpha_2 x^2+\alpha_3 x^3,$$

где коэффициенты α_i — числа, которые мы и должны отыскать, или, говоря точнее, на основании имеющихся экспериментальных данных вычислить их оценки — приближенные значения.

Конечно, можно провести целое семейство кривых — графиков подобных многочленов, которые достаточно хорошо будут соответствовать экспериментальному облаку на рис. 13. На рис. 20 нарисованы такие графики. Другими словами, можно многими способами выбрать коэффициенты многочлена так, чтобы он похоже воспроизводил ту самую неизвестную зависимость, которая, как мы надеемся, существует. Разбросы же, наглядно изображаемые облаком на рисунке, естественно считать результатом случая — их возникновение связано с погрешностями измерений, «ошибками» природы или другими многочисленными источниками помех.

Однако нас такое разнообразие не устраивает, ибо если на самом деле имеется функциональная зависимость $y=f(x)$ между x и y, то среди всех математических моделей выбранного вида (в обсуждаемом примере — среди всех многочленов третьей степени) должна же быть наиболее тесно примыкающая к функции $f(x)$. Как же

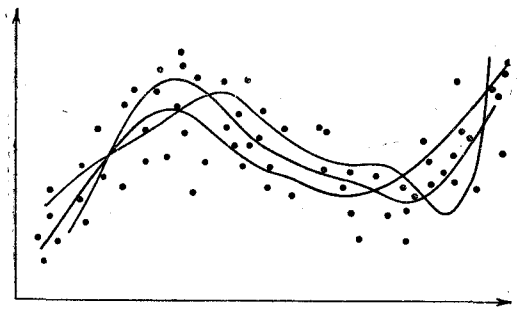


Рис. 20.

найти эту наилучшую модель среди всех моделей выбранного вида?

Читатель, искушенный в таких вопросах или хотя бы обративший внимание на раздел «Немного о критериях», вспомнит о необходимости сначала выбрать меру тесноты примыкания одной функции к другой — критерий близости двух функций.

Я не буду сейчас делать отступление и обсуждать всякие возможные меры, а сразу буду двигаться в нужном направлении и расскажу об одной весьма плодотворной идее выбора наиболее подходящей модели.

Обратите внимание: функция $f(x)$, которую мы хотим воспроизвести, хотя бы и приближенно, нам неизвестна. Вся имеющаяся о ней информация содержится в облаке экспериментальных данных и тех соображениях, на основании которых выбрана математическая модель.

Если начертить график одной из возможных реализаций математической модели выбранного нами вида, то есть при каких-то определенных числовых значениях коэффициентов, то кривая как-то пройдет среди точек экспериментального облака. Соединим теперь каждую экспериментальную точку с кривой посредством отрезка прямой,

параллельной оси ординат, как и сделано на рис. 21. Все эти отрезочки в совокупности и отражают некоторым образом, сколь хорошо начерченная кривая соответствует экспериментальным точкам.

Но отрезочков много, все они разного размера, и нужно придумать, как совместно использовать всю совокупность отрезочков, превратить эту совокупность в одно число — критерий.

На оси ординат направление вверх — положительное, так что величина отрезочка от точки до кривой будет положительной, когда точка лежит над кривой, и отрицательной, когда точка под кривой. Поэтому алгебраическая сумма отрезочков не характеризует интересующую нас величину, и нужно предложить что-то лучшее.

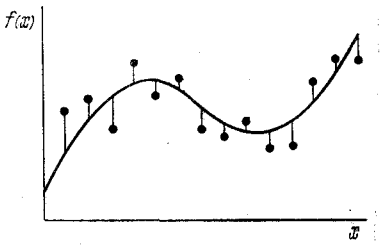


Рис. 21.

Можно взять сумму длин отрезочков, отбросив знаки, то есть сумму их абсолютных величин, но значительно более удобной мерой оказывается сумма квадратов длин отрезочков. И следует выбрать ту модель, или, говоря точнее, те значения коэффициентов α_0, α_1, α_2, α_3, при которых сумма квадратов длин отрезочков будет минимальной.

Естественно, вы хотите знать, почему я сейчас предлагаю сумму квадратов, а не сумму четвертых степеней или какую-нибудь другую функцию этих длин.

Предложил взять в качестве меры уклонения сумму квадратов все тот же Лежандр в 1805 году в статье «Новые методы определения орбит комет», где он пишет: «После того как полностью использованы все условия задачи, необходимо определить коэффициенты так, чтобы величины их ошибок были наименьшими из возможных Для этого нами указан, вообще говоря, простой метод, который состоит в отыскании минимума суммы квадратов ошибок».

Как видите, Лежандр не позаботился пояснить, на основании каких соображений он выбрал именно сумму квадратов. Однако на самом деле для такого выбора есть весьма глубокие основания. Метод выбора коэффициентов математической модели, основанной на минимизации

суммы квадратов уклонений, называется **методом наименьших квадратов**.

Вскоре великий Гаусс в ряде работ привел теоретико-вероятностное обоснование метода наименьших квадратов. Впрочем, Гаусс, который был моложе Лежандра на 25 лет, отстаивал и приоритет использования метода наименьших квадратов с 1795 года: даже великие ученые подчас проявляют мелочность...

В первых работах метод наименьших квадратов тесно связывался с нормальным распределением ошибок измерений, и для обоснования метода Гаусс воспользовался методом наибольшего правдоподобия. Но наиболее важные свойства оценок коэффициентов, получаемых по методу наименьших квадратов, как сам же Гаусс в дальнейшем показал, в действительности не зависят от распределения.

Идею доказательства Гаусса я покажу все на том же кубическом уравнении регрессии. Оценки неизвестных коэффициентов α_0, α_1, α_2, α_3 обозначим соответственно a_0, a_1, a_2, a_3. В результате наблюдений получено n пар значений (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Соответствующие точки на плоскости (x, y) и образуют наше облако. Будем считать, что значения $x_1, x_2, \ldots, x_n$ отсчитываются безошибочно: в физических, экономических или технологических задачах часто переменная x во власти исследователя — это может быть время, устанавливаемая температура или что-то подобное. Таким образом, случайные отклонения от искомой точной зависимости

$$y=\alpha_0+\alpha_1 x+\alpha_2 x^2+\alpha_3 x^3$$

для каждой точки (x_i, y_i) — это вертикальные отрезочки на рис. 21. Они представляют собой разности

$$\delta_i = y_i - (\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3).$$

Будем интерпретировать теперь отыскание оценок коэффициентов a_0, a_1, a_2, a_3 как азартную игру, в которой выиграть нельзя, а можно лишь проиграть. За меру проигрыша примем сумму квадратов $\sum \delta_i^2$, так что проигрыш тем больше, чем больше величины погрешностей δ_i, то есть размеры отрезочков.

Сформулирую теперь главные требования к оценкам. Во-первых, оценки не должны содержать систематических погрешностей, то есть математическое ожидание оценки a_k должно равняться соответствующему коэффициенту α_k. Это же означает равенство нулю математических ожиданий каждой из погрешностей δ_i. Во-вторых, математическое ожидание суммы квадратов проигрышей, то есть дисперсия суммарной ошибки, должно быть наименьшим среди всех других оценок.

Оказалось, что оценка, удовлетворяющая этим двум требованиям, как раз и приводит к оценкам коэффициентов, получаемым по методу наименьших квадратов.

Приложили свой гений к методу наименьших квадратов Лаплас, П. Л. Чебышев и А. А. Марков. Последний весьма содержательно обобщил результаты Гаусса, так что эти факты в настоящее время известны как теорема Гаусса — Маркова.

Вычисление оценок метода наименьших квадратов сводится к решению некоторой системы линейных алгебраических уравнений, принципиальных трудностей здесь нет, но есть заметные достижения в вычислительных процедурах. Однако они требуют более профессионального изложения.

Искусство надежды

Чемпион по боксу, встретив незнакомую одинокую собачку, приостанавливается: хотя боксер значительно сильнее собачки, он не хочет подвергнуться нападению, результатом которого могут быть не только порванные штаны, но и необходимость ходить на уколы. Он, конечно, надеется на лучшее, и потому не идет в атаку и не прячется в подъезд, но все же однозначно предугадать действия собачки боксер не может и выжидает — собирает необходимую информацию для предсказания ее поведения. Дрессировщица тоже надеется на лучшее, когда кладет голову в пасть льву, в то время как зрители все же имеют в виду и возможность трагического исхода — иначе им нечем было бы восхищаться. Но дрессировщица владеет искусством надежды — в противном случае лев откусил бы ей голову задолго до выхода на арену.

Об искусстве надежды великолепно написал В. Леви в интересной книжке «Охота за мыслью»: «Мы — порождение гибкого, подвижного, многомерного мира живой природы, изобилующего вариантами, переполненного возможностями и неожиданностями. В этом мире мало на что можно надеяться твердо, но именно поэтому живым существам пришлось учиться искусству надежды. Это была суровая школа. Те, кто плохо надеялся, умирал раньше других.

Хорошо надеяться — это значит надеяться не переставая. Но не жестко и тупо, а гибко. Не слепо, а зорко и точно. Хорошо надеяться — это значит хорошо выбирать, на что можно надеяться. Это значит уметь вовремя переменить выбор. Это значит уметь взвешивать шансы, оценивать вероятности. Короче, уметь предвидеть и полагаться на предвидение.

Искусство надежды — это искусство достижения цели. Цели могут быть самые разные.

Искусством надежды владеет и ласточка, летящая наперерез мошке, и кошка, стерегущая мышь. Этим искусством владеет вратарь, караулящий мяч в нужном углу ворот, и стрелок, метящий в подвижную цель».

Когда девушка гадает на ромашке «любит, не любит, плюнет, поцелует...» или слушает бормотание цыганки: «по дальней дороге — казенный дом, по ближней дороге — получишь письмо...» и, опираясь на это, строит свою надежду на интимные успехи, то она совсем не владеет искусством надежды.

Все суеверия построены по тому же принципу: надежда опирается на случайные совпадения, на факты, не связанные причинно-следственными связями. Приведу анекдотический пример из современной действительности, на который обратил мое внимание В. М. Глоговский.

За день до финала розыгрыша Кубка СССР по футболу в газете «Советский спорт» 3 сентября 1976 года в статье К. Есенина «Что нас ждет в тридцать пятом?» написано: «Я всегда люблю искать какие-то штрихи, которые могут что-то подсказать по поводу исхода предстоящего финала. Так вот, высказываю предположение, что победит тот, кто первым забьет гол. По крайней мере, в последнее время была замечена такая закономерность: в нечетные годы выигрывал тот, кто первый пропускал гол в свои ворота, а в четные тот, кто первым забивал».

А на следующий день состоялась финальная игра, где встретились команды «Динамо» (Тбилиси) и «Арарат». Первыми забили гол тбилисцы... и затем выиграли кубок. Год был четный и, таким образом, гипотеза К. Есенина еще раз подтвердилась. Но, думаю, автор статьи, когда выдвигал эту гипотезу, понимал ее несерьезный характер, и написал все это в шутку, ибо здесь совершенно очевидно полное отсутствие причинно-следственных связей между результатами или ходом игры и четностью года. Поэтому, если рассматривать подобную гипотезу всерьез, то

следует ее отмести как абсолютно бессодержательную и влекущую за собой лишь предрассудки.

И все же постоянно кто-то гадает на монете или на чем-то аналогичном, и выдает подобное гадание за нечто серьезное. Однако выпади герб сейчас или нет, в модели с равновозможными и независимыми исходами шансы выпадения герба при следующем розыгрыше те же, что и раньше. Если же герб выпал подряд несколько раз, то у игрока возникает недоверие к самой модели, к предположениям о равновероятности и независимости, хотя такая ситуация принципиально возможна.

Подобный эксперимент я проводил неоднократно со студентами: если при подбрасывании монеты, кости или при другом эксперименте с исходами, вероятности которых априори одинаковы, семь или десять раз подряд выпадал один из них, то почти вся аудитория при голосовании поднимала руки за пересмотр гипотезы о равновозможности исходов, и чаще всего подозревала меня в надувательстве, и притом совершенно справедливо.

В действительности последовательность независимых опытов не несет никакой информации о будущем при равновероятности исходов и доставляет довольно бедную информацию при неравных вероятностях, причем, конечно, несет тем большую информацию, чем больше различаются вероятности исходов. Все наши рассуждения призывают при игре в монетку, кости или рулетку, не забывая о теории вероятностей, опираться на здравый смысл.

Но эта книжка — не сборник полезных советов по поведению в игорном доме. А в реальной жизни последующие события в значительной степени определяются предыдущими, и именно на эту зависимость опирается искусство надежды или, говоря другими словами, предсказание, прогноз.

На обложке вы видите игральную кость со словами «ДА, НЕТ, МОЖЕТ БЫТЬ». Однако этот образ лишь частично отражает содержание книги. Игральная кость, конечно, имеет прямое отношение к понятию вероятности — она обычно используется как модель событий с равновероятными и независимыми исходами. Но для научного предсказания будущего или, в более широком плане, для управления в условиях статистической устойчивости необходимо использовать сведения о прошлом. Решения здесь не могут опираться на результаты подбрасывания кости: фатализм — это альтернатива искусству надежды, и поэтому

образ игральной кости не отвечает линии содержания, касающейся проблем научного прогноза. Но художник К. Р. Борисов не согласился с моими сомнениями, ссылаясь на мою же позицию: обложка — модель книги, а модель может описывать, отражать объект лишь с какого-то ракурса... Мне пришлось сдаться, а уж, вы читатель, решайте, кто из нас прав.

Законы Ньютона дали возможность точно предсказывать движение планет. Отклонения силы тяготения от известного всем закона совершенно ничтожны. Движение снаряда при стрельбе из пушки тоже может быть предсказано довольно точно, но все же со значительно меньшей точностью, чем движение планет. Здесь сказывается влияние множества причин, искажающих траекторию движения снаряда по сравнению с расчетной: разбросы веса снарядов и взрывчатого вещества, искажения геометрических форм снаряда и ствола орудия, неоднородности атмосферы, поля силы тяжести Земли и т. д. В результате, как вам известно, снаряды нечасто точно попадают в цель, они ложатся вблизи цели, наблюдается разброс... Но я сейчас хочу подчеркнуть, что траектория снаряда может быть предсказана довольно точно, и при владении искусством стрельбы попадание в неподвижную цель — задача, практически решенная довольно давно.

Совсем иначе выглядит задача попадания в движущуюся цель.

Вы, конечно, знаете, как это трудно — попасть в подвижную цель. Если вы не охотились, то, небось, играли в лапту, волейбол или баскетбол, и представляете себе трудности: ведь нужно попасть не в то место, где сейчас находится цель, а туда, где она будет, когда мяч долетит. Но вы же точно не знаете, как будет двигаться цель, будь то ваш противник или партнер, и целитесь в то место, где, как вы надеетесь, как раз и окажется цель, когда мяч долетит...

Множество промахов в вашей жизни в прямом и переносном смысле показывает, к сожалению, что искусство достижения движущейся, изменяющийся цели — искусство надежды дается не легко, и даже владение этим искусством не избавляет вас полностью от промахов. Но не будем сейчас отвлекаться от простой и четкой задачи попадания в движущуюся цель.

В Первую мировую войну стреляли по «допотопным» самолетам из винтовок или небольших пушек. Но ко Второй мировой войне самолеты поднялись так высоко и ско-

рости их столь возросли, что нечего было всерьез надеяться попасть в самолет из винтовки, хотя все же такие случаи были. На всякий яд есть противоядие — появилась зенитная артиллерия. Почти невозможно попасть «на глазок» в самолет, движущийся со скоростью в несколько сот километров в час, да еще и маневрирующий, дабы вас запутать — человеческие возможности здесь ограничены. Поэтому в период перед второй мировой войной и во время войны возникла проблема создания автоматических устройств для наведения зенитных орудий на цель.

В 1948 году вышла знаменитая книга Норберта Винера «Кибернетика или управление и связь в животном и машине», перевернувшая мир наших представлений об управлении. Одной из задач, на которые Винер опирался при построении общих положений кибернетики, была задача об управлении огнем зенитной артиллерии. Я сейчас приведу схему рассуждений Винера, ибо они послужили основой очень многих последующих работ по управлению в ситуации, где необходим прогноз поведения изучаемого объекта. Опубликованы они в другой его книге, появившейся вскоре после выхода «Кибернетики» *).

Если бы самолет двигался равномерно и прямолинейно, то, определив с помощью радиолокатора его местоположение и вектор скорости, можно было бы легко предсказать ту точку, где он будет через время τ, необходимое зенитному снаряду для преодоления расстояния от орудия до самолета, и нацелить орудие именно в эту точку. Но самолет не движется равномерно и прямолинейно даже в течение небольшого интервала времени τ: его траектория, с позиций стороннего наблюдателя, случайна, однозначно не предсказуема.

Однако изучение записей траекторий позволяет построить математическую модель траектории движения самолета, и в качестве такой модели Винером был выбран стационарный случайный процесс. Стационарность, грубо говоря, означает однородность во времени вероятностных характеристик процесса.

Вспомним о кирпичиках, из которых можно сложить нужные функции. С таких позиций удобно представлять себе стационарный случайный процесс как сумму гармонических колебаний различных частот (не обязательно

*) N. Wiener, Extrapolation, Interpolation and Smoothing of Stationare Time Series, John Wiley, New York, 1949.

кратных какой-то основной частоте), амплитуды которых — независимые случайные величины. Каждый стационарный случайный процесс с любой степенью точности может быть приближен подобной тригонометрической суммой.

Теперь поясню, какую же надо решить задачу. Устройство, управляющее огнем зенитного орудия, должно определить направление на точку встречи снаряда и цели и дать команду на выстрел. Определение точки встречи затруднено тем, что траектория цели случайна — каждая из координат (например, дальность до цели и азимутальные углы) есть одна из реализаций стационарного случайного процесса. Займемся какой-либо одной из координат. Обозначим

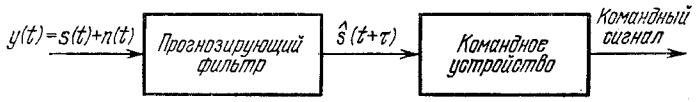


Рис. 22.

сигнал, поступающий с радиолокатора на управляющее устройство, через $y(t)$; он складывается из сигнала, соответствующего наблюдаемой координате траектории цели $s(t)$,— как говорят, полезного сигнала,— и помехи $n(t)$. Помехи всегда присутствуют — это собственные шумы аппаратуры, атмосферные разряды и т. д. Такие помехи с достаточной степенью точности также можно считать реализациями некоторого стационарного случайного процесса, конечно, отличного от полезного сигнала и статистически независимого от него. Таким образом, на управляющее устройство (рис. 22) поступает смесь сигнала и шума:

$$y(t) = s(t) + n(t).$$

Задача первого блока — прогнозирующего фильтра — выработать сигнал $\hat{s}(t+\tau)$, возможно более близкий к реальному значению сигнала $s(t+\tau)$ в момент $t+\tau$ встречи снаряда и цели.

Иными словами, нужно выработать сигнал $\hat{s}(t+\tau)$, дающий наилучшее приближение к реальным координатам цели через время τ, необходимое снаряду для достижения цели.

Я хочу подчеркнуть особенности описанного подхода. Если бы сигнал $s(t)$ (а вместе с ним и траектория цели) был известен точно, то принципиально задача не представ-

ляла бы никаких трудностей — она сводилась бы к арифметическим вычислениям. Если бы шум $n(t)$ был известен точно, то, так как наблюдаемый сигнал $y(t)$ известен, а $s(t) = y(t) - n(t)$, то и $s(t)$ определялся бы точно, и задача опять была бы тривиальна. В другой крайней ситуации, когда априори никакой информации относительно $s(t)$ и $n(t)$ нет, то нет и никакой надежды построить хорошую оценку, опираясь лишь на смесь $s(t) + n(t)$. Последняя ситуация аналогична подбрасыванию игральной кости — будущее никак не определяется прошлым. Однако реально инженеры многое знают и о структуре помех, и о возможных траекториях самолетов или других движущихся объектов, и именно эти знания дают возможность строить математическую модель, о которой идет речь.

Теперь, конечно, нужно уточнить критерий качества прогноза или критерий близости действительного значения сигнала $s(t+\tau)$ и прогнозируемой фильтром оценки $\hat{s}(t+\tau)$. Как мы уже неоднократно обсуждали, критериев можно предложить сколько угодно. Но, следуя логике метода наименьших квадратов, разумно взять в качестве критерия близости $s(t+\tau)$ и $\hat{s}(t+\tau)$ математическое ожидание квадрата их разности:

$$\rho = \mathsf{M}\,[s(t+\tau) - \hat{s}\,(t+\tau)]^2.$$

Тогда задача выбора наилучшего прогноза сводится к выбору фильтра, минимизирующего величину ρ. Следует, конечно, заметить: такой фильтр не будет давать наилучший прогноз всегда — он указывает значения $\hat{s}(t+\tau)$, которые лишь в среднем дают наилучший прогноз. Но мы находимся в условиях случайности, и ничего лучшего, обеспечивающего оптимальность прогноза всегда, предложить нельзя.

Решение задачи о наилучшем прогнозе поведения стационарного случайного процесса имеет наглядную геометрическую интерпретацию.

Общее понятие линейного векторного пространства дает возможность рассматривать совокупность случайных величин (будем обозначать их греческими строчными буквами), математическое ожидание которых равно нулю, а дисперсия ограничена, как гильбертово пространство, причем скалярным произведением случайных величин — векторов нашего пространства — служит математическое

ожидание их произведения

$$(\xi, \eta) = \mathsf{M}\,\xi\eta,$$

а квадрат длины вектора — это его дисперсия. Расстоянием между двумя векторами ξ и η тогда будет корень квадратный из дисперсии разности векторов:

$$\|\xi - \eta\| = \sqrt{\mathsf{M}\,(\xi - \eta)^2}.$$

Пусть $\xi(t)$ — случайный процесс. В каждый момент времени t случайной величине $\xi(t)$ отвечает вектор в гильбертовом пространстве H случайных величин. При изменении времени t меняется и случайная величина $\xi(t)$ — ее вектор переходит в другое положение. Когда время t пробегает какой-то отрезок $a \leqslant t \leqslant b$, то конец вектора в пространстве H описывает некоторую кривую, и на множество всех ее векторов можно натянуть подпространство; обозначим его $H\,(a, b)$.

Пусть теперь t — момент наблюдения, а нас интересует значение случайного процесса в будущий момент $t+\tau$. Вообще говоря, предсказать будущие значения $\xi\ (t+\tau)$ однозначно нельзя, и лучшее, что возможно сделать, это указать оптимальную оценку этих значений. Геометрическая интерпретация сразу указывает решение. Построим подпространство $H(a, t)$ — оно отвечает прошлым (до момента t) значениям случайного процесса. Если $\tau > 0$, то вектор $\xi\ (t+\tau)$ — «будущее» процесса $\xi(t)$ в момент $t+\tau$, вообще говоря, не содержится в «прошлом» подпространстве $H(a, t)$; в противном случае был бы возможен точный прогноз. Наилучшим приближением будущего значения $\xi(t+\tau)$ *по прошлому*, то есть по наблюдениям над $\xi(t)$ в интервале (a, t) будет проекция вектора $\xi(t+\tau)$ на прошлое подпространство $H(a, t)$, а длина перпендикуляра из конца вектора $\xi(t+\tau)$ на подпространство $H(a, t)$ будет равна ошибке прогноза.

Если процесс $\xi(t)$ — стационарный и обладающий некоторыми свойствами, обычно встречаемыми в практических задачах, то эти геометрические соображения дают возможность написать вычислимые выражения для прогноза и даже построить устройства — прогнозирующие фильтры, реализующие наилучший прогноз.

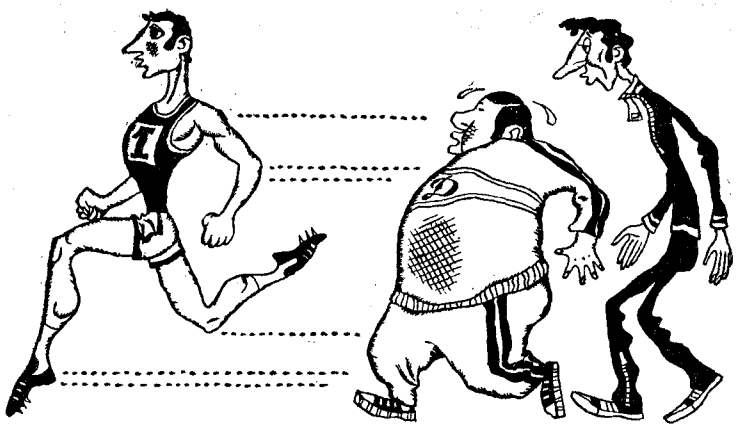
Математический метод определения характеристик наилучшего линейного прогнозирующего фильтра, разработанный Винером, достаточно сложен и опирается на

развитые математические методы функционального анализа, интегральных уравнений, функций комплексного переменного. Поэтому я ничего о нем здесь рассказывать не буду. Хочу лишь заметить, что методы предсказания в теории стационарных случайных процессов были разработаны в 1939—1941 годах знаменитым советским математиком академиком А. Н. Колмогоровым, однако Н. Винер, работавший в этом направлении независимо, не только создал соответствующую теорию, но и применил ее к очень важной технической проблеме. Развитию, уточнению и различным применениям теории предсказания случайных процессов посвящена огромная литература. Здесь и прогноз погоды, и управление автопилотом, системы слепой посадки самолетов, прогноз спроса и поставок в материально-техническом снабжении, прогнозирование экономических показателей при текущем и перспективном планировании и многое другое.

Методы, идущие от Винера и Колмогорова, дают возможность изучить не только задачи прогнозирования, но и целый пласт проблем радиофизики и радиотехники, физики атмосферы, теории регулирования и т. д. Так бывает в науке часто: для изучения какой-то проблемы в физике, технике, биологии или в других разделах науки разрабатывается математический аппарат, а затем открываются возможности использования этого аппарата для решения других задач в тех же, а то и иных разделах науки.

В связи с задачами статистической механики замечательный советский математик А. Я. Хинчин выделил класс стационарных случайных процессов и разработал математический аппарат для его изучения — это было в 1935 году *). А в сороковых годах этот математический аппарат был эффективно применен в радиотехнике для решения задачи фильтрации — отделения передаваемого сигнала от шума. Вот простейшая модель. На вход приемного устройства поступает детерминированный полезный сигнал $s(t)$ и шум $n(t)$ — стационарный случайный процесс. Приемное устройство должно выделить сигнал и, по возможности, подавить шум. Задача теории на первых порах состояла в вычислении отношения мощностей сигнала и шума на выходе приемного устройства. Впоследствии задачи все усложнялись — нужно было выбирать форму сигнала и характеристики приемного устройства, дающие возможно большее отношение сигнала к шуму на выходе приемника. Дальнейшее развитие теории привело к более общим постановкам: нужно выбрать так характеристики приемника, чтобы оптимизировать в смысле некоторого критерия соотношение характеристик сигнала и шума на выходе приемника. При этом не только шум $n(t)$, но и сигнал $s(t)$ рассматриваются как стационарные случайные процессы. Ситуацию, когда $s(t)$ — детерминированный сигнал, можно рассматривать как частный случай стационарного процесса. Вот здесь уже можно воспользоваться теорией Винера — Колмогорова, считая, что время прогноза $\tau=0$, и задача состоит в выборе такой конструкции приемника (в математической постановке — выбора такого оператора), при котором сигнал на выходе будет, в смысле метода наименьших квадратов, отличаться от полезного сигнала на входе как можно меньше, то есть когда математическое ожидание квадрата уклонения сигнала на выходе от полезного сигнала на входе будет минимальным. Эта задача решается теми же методами, что и задача прогноза, о которой шла речь выше.

*) А. Я. Хинчин, Теория корреляции стационарных стохастических процессов, Успехи математических наук, вып. 5, 1938.



В борьбе за рекорд

В предыдущем разделе речь шла о прогнозе будущих значений координат движущегося объекта. Сейчас мы обсудим другие задачи предсказания.

Классный спортсмен — бегун на короткие дистанции собирается выступить в соревнованиях. Его личный рекорд — 10,4 сек при беге на 100 метров, но ему хочется сбросить хотя бы две десятки: при времени 10,2 сек он по-видимому, добьется в ближайших соревнованиях первенства... Но эти десятые доли секунды даются с большим трудом, и, следовательно, нужен какой-то другой режим. Ясно, что время, затрачиваемое на дистанцию, зависит и от продолжительности тренировок, и от режима питания. Можно изменить — увеличить или уменьшить — длительность тренировок, можно изменить режим питания, увеличив или уменьшив число ежесуточно потребляемых калорий. Как же следует поступить? С одной стороны, бегун должен иметь возможно меньший вес, и, следовательно, его нужно весьма умеренно кормить. С другой стороны, он должен быть физически сильным, чего при голодовке не добьешься,— для развития мышц нужно интенсивное питание. Таким образом, продолжительность бега — функция от количества ежесуточно потребляемых спортсменом калорий, причем как слишком большое, так и слишком малое их количество приводит к увеличению продолжительности пробега. Похоже, что

график зависимости времени пробега стометровки от количества потребляемых калорий имеет вид кривой, подобной параболе или другой функции, имеющий один минимум.

Если спортсмен тратит на тренировки мало времени, то его мышцы будут слабыми, нужные рефлексы — притупленными, и его результаты будут плохими. Но если он тренируется по пятнадцать часов в сутки без отдыха, то также нельзя ожидать ничего хорошего — организм устанет, наступит вялость, безразличие — какие уж здесь рекорды... Поэтому время пробега стометровки как функция продолжительности ежедневных тренировок выглядит, по-видимому, также подобно параболе.

Нужно подчеркнуть: из наших рассуждений следует наличие наилучшего режима тренировок, обеспечивающего индивидуальный рекорд, и теперь нужно указать способ его отыскания. Конечно, режим индивидуален, общей для всех спортсменов формулой его не задашь, и нужно предложить способ определения оптимального режима, пользуясь лишь наблюдениями, данными эксперимента.

Сейчас мы воспользуемся для решения задачи идеями регрессионного анализа. Время τ пробега стометровки есть по нашей гипотезе функция двух переменных или факторов — времени ежесуточных тренировок t и ежесуточкого количества потребляемых спортсменом калорий v:

$$\tau = f(t, v).$$

Вид этой функции неизвестен, но общие соображения дают возможность предположить, что она представляет собой параболоид, уравнение которого имеет вид

$$\tau = \beta_0 + \beta_1 t + \beta_2 v + \beta_{11} t^2 + \beta_{12} tv + \beta_{22} v^2. \qquad (\beta)$$

Это и есть уравнение регрессии, отличающееся от рассмотренных выше лишь тем, что здесь не одна, а две независимые переменные t и v.

Если бы были известны коэффициенты β_0, β_1, ..., β_{22}, то легко можно было бы найти режим, соответствующий наименьшему времени пробега, то есть определить значения факторов t и v, отвечающие индивидуальному рекорду — минимальному времени пробега стометровки. Другими словами, можно было бы определить координаты t_0, v_0 самой низкой точки τ_0 параболоида. Однако коэффициенты неизвестны, и спортсмен и его тренер могут лишь проводить наблюдения, на основании которых можно будет

найти приближенные оценки для коэффициентов и соответственно приближенные значения для оптимальных параметров t_0, v_0.

Каждый эксперимент длится довольно значительное время: потребуется подчас не одна неделя для перестройки организма на новый режим и получения тех результатов, которые этот режим обеспечивает. Поэтому уж очень много режимов не перепробуешь, а хочется найти оптимальный. Кроме того,— и это очень важно для понимания дальнейших рассуждений — при выбранном режиме питания и тренировок результат в беге неоднозначен: то спортсмен недоспал, то переволновался из-за каких-то личных дел, то погода была неудачная. Словом, по многим причинам, не учтенным в выбранной математической модели, наблюдается разброс результатов пробега стометровки при одном и том же режиме. Мы будем полагать этот m разброс случайным, и поэтому будем пользоваться при оценке коэффициентов модели статистическими методами. Впрочем, вы уже, наверное давно догадались, куда я клоню: нужно для оценки коэффициентов воспользоваться методом наименьших квадратов. В предыдущих разделах метод наименьших квадратов был использован для функций одного переменного, но наличие нескольких переменных почти ничего не меняет ни в идейной части, ни в формализме: для получения оценок коэффициентов методом наименьших квадратов нужно решить систему линейных алгебраических уравнений относительно неизвестных коэффициентов β_0, β_1, ..., β_{22}. А известными при этом будут результаты наблюдений: при каждом выбранном режиме (t_i, v_i) спортсмен пробежит стометровку несколько раз и покажет результаты $\tau_i^{(1)}$, $\tau_i^{(2)}$, ..., $\tau_i^{(m)}$, вообще говоря, различные.

Геометрическая картина выглядит так: на плоскости отмечаются точки — опробованные режимы, и над каждой из них расположена точка, вертикальная координата которой равна достигнутому результату.

Теперь для решения задачи нужно в пространстве построить такую поверхность параболоида — математическую модель, которая лучше всего соответствовала бы полученным точкам (t, v, τ), а уже затем на этой поверхности нужно найти минимум τ_0 и его координаты (t_0, v_0), которые и будут взяты в качестве наилучшего режима.

Внимательный читатель заметит, что на самом деле здесь обсуждалось сразу две проблемы: идентификация — построение модели по данным наблюдений, и оптимизация — выбор наилучшего режима. Давайте разделим эти проблемы и посмотрим, что же мы сделали для решения каждой из них и какие следуют выводы из проведенного обсуждения. Начну с проблемы идентификации.

Нам известны достижения спортсмена при нескольких наблюденных режимах. Спортсмена и тренера интересует, конечно, вечный вопрос: «Что было бы, если бы?...» — в данном случае если бы они выбрали другие режимы, например установили режим питания в 4800 калорий при продолжительности тренировок в 5 часов. Если математическая модель достаточно хорошо представляет имеющиеся экспериментальные данные, как говорят, модель адекватна, то с ее помощью можно ответить на поставленные вопросы. Для этого нужно лишь в формулу

$$\tau = b_0 + b_1 t + b_2 v + b_{11} t^2 + b_{12} tv + b_{22} v^2 \qquad \text{(b)}$$

подставить значения $\bar{t}$ и $\bar{v}$, соответствующие интересующему их режиму, и после простых вычислений получить значение $\tau = \tau(\bar{t}, \bar{v})$. Черточки над t и v я поставил, чтобы подчеркнуть: значения $\bar{t}$ и $\bar{v}$ — фиксированные, это же выбранный режим: $\bar{t}=5$ часов, $\bar{v}=4800$ калорий.

Заметьте: в двух последних формулах, обозначенных (β) и (b) и имеющих совершенно одинаковый вид, коэффициенты обозначены разными, но похожими буквами. Это, конечно, неспроста: в первой формуле коэффициенты представляют собой определенные числа, а в формуле (b) подставлены коэффициенты, найденные по методу наименьших квадратов, и они являются оценками соответствующих коэффициентов в формуле (β). Такая система обозначений широко распространена в регрессионном анализе.

Но если вместо точных значений коэффициентов в формулу подставлены их оценки, то и вместо значения функции τ будет найдена лишь ее оценка, ее приближенное значение. При этом оценка будет тем менее точной, чем по меньшему количеству экспериментальных точек она построена.

Итак, по экспериментальным данным можно построить уравнение регрессии (b) и с его помощью осуществлять предсказание значений интересующего исследователя параметра (в данном случае времени τ пробега стометровки) в точках (режимах), расположенных внутри обследуемой области, причем точность предсказания будет, вообще

говоря, тем выше, чем большее количество экспериментальных точек было использовано при построении уравнения регрессии.

Как вы думаете, зависит ли точность предсказания от расположения точек, в которых уже проведены наблюдения? Ответ здесь не очевиден и требует некоторых размышлений...

Теперь я хочу заметить: при обсуждении режима спортсмена мы выбрали лишь два фактора — продолжительность тренировок и количество калорий. Но, конечно, состояние спортсмена, его возможности зависят не только от этих факторов. Скажем, если выбрана продолжительность тренировок 6 часов, то можно их провести с девяти утра до трех часов дня без перерыва, можно разбить на две тренировки по три часа или на три тренировки по два часа, и, по-видимому, результаты будут разными.

Кроме того, и сами тренировки могут быть весьма разнообразными: каждый спортсмен должен развиваться гармонически, так что в тренировки бегуна включаются и другие виды легкой атлетики, и гимнастика, и тяжелая атлетика и т. д.

В питании спортсмена не все решают калории: ему нужен разнообразный стол, где есть белки, жиры, углеводы, витамины, и их количество можно варьировать по-разному. Наконец, кроме тренировок и питания, еще многое определяет состояние спортсмена: продолжительность сна (далеко не всегда лучше переспать, чем недоесть!) возраст, да и многие другие факторы, хотя бы и не выражаемые в числах, но весьма важные, такие, как моральное состояние или общий культурный уровень.

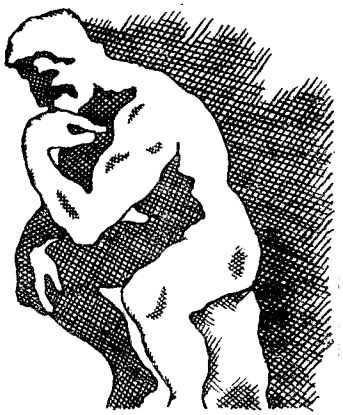
Таким образом, режим жизни и общее состояние спортсмена описывается довольно большим числом факторов, и задача выбора оптимального режима представляется весьма не простой.

Оставим в покое на некоторое время режим спортсмена и обратимся к проблеме выбора оптимального режима для действующей заводской технологической установки.

Здесь наблюдаемыми факторами будут технологические параметры процесса, например температуры, давления, расходы, концентрации веществ, участвующих в физико-химических реакциях. В режиме нормальной эксплуатации их значения не задашь по своему усмотрению — они такие, какие есть, вмешиваться в действующую установку трудно по многим причинам.

Пример такой ситуации был подробно рассмотрен — это процесс электрообессоливания. Но выше мы обсуждали лишь взаимосвязь входа и выхода, более точно — количества солей на входе и выходе ЭЛОУ. Однако в действительности эффективность процесса обессоливания зависит от многих факторов, скажем, от температуры сырья, количества промывочной воды и добавляемого деэмульгатора, величины напряженности электрического поля и времени пребывания эмульсии в электрическом поле.

Если факторов больше двух, то наглядные геометрические представления отступают, хотя, как я уже говорил, и ассоциации, и терминология сохраняются: мы говорим о многомерном факторном пространстве и поверхности в этом пространстве, отражающей зависимость интересующего нас параметра (например, количества солей на выходе ЭЛОУ или времени пробега стометровки) от всех учитываемых факторов. Эта поверхность в теории эксперимента называется поверхностью отклика; по-видимому, имеется в виду отклик изучаемого объекта на воздействие в виде некоторого набора факторов. И если речь идет об идентификации, то задача исследователя — построить уравнение поверхности отклика (оно называется функцией отклика) и проверить, сколь хорошо выбрано уравнение — проверить адекватность. И если адекватности нет, то надо все начинать сначала: угадывать или, говоря более научно, выбирать другую математическую модель, вновь проверять адекватность и т. д., пока не убедишься в адекватности,— вот тогда можно уже воспользоваться полученным уравнением для предсказания значений отклика в точках, расположенных в обследуемой области и, конечно, отличных от тех, по которым строились оценки коэффициентов регрессии.



Пороки пассивности

Наблюдение за движением кометы, за течением технологического процесса в режиме нормальной эксплуатации, за достижениями спортсмена, когда его режимы выбраны «как попало»,— все это примеры пассивных экспериментов.

Недеятельного, инертного, безучастного, безразличного к окружающей жизни человека называют пассивным, и часто в тоне звучит осуждение. Но не всегда пассивность — вина исследователя. Астроном, наблюдая за движением кометы, может лишь «безучастно» регистрировать ее координаты или красочно описывать видимые эффекты: на ее движение он повлиять никак не может, во всяком случае, пока еще не может...

При наблюдении за течением технологического процесса на действующей заводской установке в режиме нормальной эксплуатации технолог или оператор может и должен изменять параметры процесса, но эти изменения определяются регламентом и весьма незначительны. На многих установках действуют автоматические регуляторы — они поддерживают практически постоянными регулируемые параметры — давления, температуры, расходы, напряжения, токи, скорости и многое другое. В этих условиях исследователь, изучающий процесс, не имеет возможности вмешиваться и заметно увеличивать или уменьшать интересующие его параметры — ему отводится роль пас-

сивного наблюдателя: регистрируй, записывай, что угодно, но «нам» не мешай. И в самом деле, всякое «постороннее» вмешательство может привести к разладке регулируемого процесса, качество продукции может при этом снизиться, ее количество уменьшиться, и в результате план не будет выполнен. Этого, с позиций завода, допустить нельзя! Конечно, не исключена возможность и улучшения качественных показателей процесса, его интенсификация, повышение эффективности, но заводского работника обычно в первую очередь беспокоит сегодняшний план, за невыполнение которого он несет ответственность, а забота об улучшении... ему «лучше иметь синицу в руках, чем журавля в небе».

Когда режимы жизни спортсмена выбраны бездумно — нагрузка то большая, то маленькая, питание побольше или поменьше — и эта информация используется для выбора оптимального режима, то такие эксперименты и будут пассивными.

Опираясь на данные пассивного эксперимента, можно построить уравнения регрессии и, как мы установили в предыдущем разделе, возможно предсказание значений функций отклика по уравнению регрессии *внутри* обследуемой области. Восстановление значений функции внутри обследуемой области по отдельным данным, взятым в некоторых точках области, называется *интерполяцией*. Вы, наверное, с этим термином встречались. Например, при работе с таблицами логарифмов или тригонометрических функций, когда нужно получить значение функции в точке, не имеющееся в таблице, пользуются линейной или квадратичной интерполяцией. Но весьма часто исследователю нужно знать поведение функции отклика как раз вне обследуемой области, *экстраполировать* значения функции. Можно ли воспользоваться уравнением регрессии для решения задачи экстраполяции?

В 1676 году Роберт Гук объявил об открытии закона, устанавливающего связь между удлинением пружины при растяжении и действующей (растягивающей) силой.

Методика эксперимента Гука весьма проста: к подвешенной пружине прилагается снизу сила (вес гирьки) и измеряется удлинение (рис. 23). Конечно, опыт проводится последовательно: подвешиваются все более тяжелые гирьки, и регистрируются результаты. Если по горизон-

тальной оси откладывать удлинение, а по вертикальной — нагрузки, то экспериментальные точки тесно ложатся вдоль прямой (рис. 24). Это дает серьезные основания интерполировать поведение функции отклика (зависимость удлинения от нагрузки) между экспериментально полученными точками посредством линейной зависимости, то есть полагать наличие прямой пропорциональности между

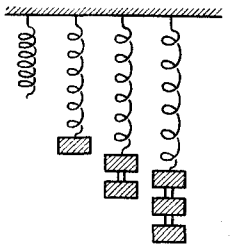


Рис. 23.

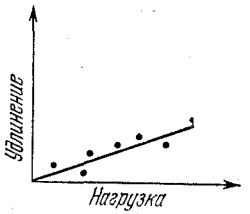


Рис. 24.

удлинением и нагрузкой внутри обследованной области нагрузок.

Но можно ли предсказать поведение функции отклика и дальше, вне участка нагрузок, проверенных экспериментально, продолжив прямую линию? Пожалуй, сомнительна такая возможность. И в самом деле, как показывают дальнейшие эксперименты, начиная с некоторых значений нагрузки, линейность нарушается, а затем нарушается и упругость. Впрочем, думаю, график (рис. 25) вам известен из курса физики, и я лишь воспользовался им для иллюстрации рассуждений о небезопасности бездумной экстраполяции.

Давайте вернемся к разделу «Идентификация технических объектов» и вспомним задачу об обессоливании нефти. На рис. 12 воспроизведены данные о количестве солей на входе и выходе ЭЛОУ, они расположены на интервале от 100 до 1000 мг/л. На всем этом участке уравнение регрессии представляет собой прямую линию. А что будет левее точки 100 мг/л? Можно ли прогнозировать ее поведение на интервале от 0 до 100 мг/л, использовав то же уравнение прямой, то есть продолжив его влево до оси ординат?

Мы уже обсуждали этот вопрос. Конечно, нельзя: если так продолжить прямую влево, то она укажет значение

15 мг/л на выходе при отсутствии солей на входе. Получается результат физически бессмысленный. Качественно ясно, что график должен проходить через начало координат и, таким образом, вид графика должен быть примерно таким, как показано на рис. 12 пунктиром. Следовательно, и здесь оказывается ошибочным предсказание зависимости вход — выход, то есть вида функции отклика в области, где нет экспериментальных точек, с помощью уравнения регрессии, построенного по данным пассивного эксперимента.

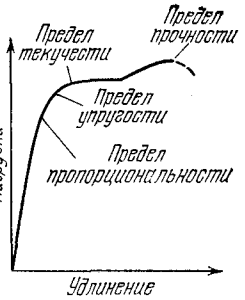


Рис. 25.

Обратите внимание на свой домашний холодильник. При склонности к исследовательской деятельности, вы можете заинтересоваться зависимостью температуры внутри холодильника от температуры помещения, числа открываний двери, количества продуктов, находящихся внутри него или от каких-либо других факторов, подверженных случайным изменениям. Вы ставите внутрь холодильника термометр и начинаете тщательно записывать все эти данные. И к своему удивлению обнаруживаете: как бы ни изменялись интересующие вас переменные (температура помещения, количество продуктов и т. д.), температура внутри холодильника — сейчас это функция отклика — меняется в малом интервале от плюс одного до плюс двух градусов, и эти изменения даже трудно обнаружить с помощью того домашнего термометра, который оказался у вас под рукой. Таким образом, несмотря на значительные изменения факторов, изменения функции отклика оказались ничтожными, сравнимыми с ошибками измерений. Это, конечно, замечательно: холодильник хорошо отрегулирован, но ваша исследовательская деятельность пошла впустую.

Если вместо холодильника вы изучаете сложный технологический процесс и обнаруживаете ту же картину практической независимости выходного параметра — функции отклика при изменениях входных переменных — факторов, то никакой содержательной математической модели процесса на основании данных пассивного эксперимента построить нельзя. Как правило, подобная ситуация означает, что процесс отрегулирован так, чтобы не реаги-

ровать на допустимые изменения входных факторов. Здесь данные пассивного эксперимента оказываются собранными в малой окрестности одного значения функции отклика, и никакую задачу предсказания, будь то интерполяция или экстраполяция, решить нельзя.

Обсудим еще некоторые стороны пассивного эксперимента, ограничивающие возможности его содержательного использования. Факторы могут оказаться как-то связанными, коррелированными. Например, при более подробном изучении зависимости времени пробега стометровки от режима спортсмена обнаруживается зависимость, скажем, между продолжительностью тренировок и сна, потребляемой жидкостью и общим объемом пищи и тому подобное. Такое положение может очень испортить дело. При вычислении коэффициентов уравнения регрессии по методу наименьших квадратов нужно решить систему линейных алгебраических уравнений. В свою очередь ее коэффициенты выражаются через значения факторов и вследствие их зависимости, будь то зависимость детерминированная или статистическая, матрица системы может оказаться плохо обусловленной. О возникающих здесь неприятностях я уже упоминал в разделе «Осторожно: задача свелась к линейной».

Неприятности возникают здесь и вследствие зависимости оценок коэффициентов регрессии. Это я поясню подробнее.

Представим себе простое уравнение регрессии, содержащее всего два фактора x_1 и x_2:

$$y=0{,}2x_1-10x_2. \qquad (*)$$

Увеличение значений фактора x_1 приводит здесь к увеличению значений отклика y. Но увеличение фактора x_2 ведет к уменьшению y. Таким образом, знак коэффициента при факторе указывает направление изменения функции отклика при увеличении фактора. Величина коэффициентов в этом уравнении указывает, конечно, сколь быстро будет изменяться функция отклика при изменении соответствующего фактора, и в уравнении (*) фактор x_2 в пятьдесят раз «более влиятелен» чем x_1.

Таким образом, абсолютные значения коэффициентов указывают относительный вклад, вносимый факторами в функцию отклика. И если бы, скажем, областью изменения факторов был единичный квадрат $0\leqslant x_1\leqslant 1$, $0\leqslant x_2\leqslant 1$, то уравнение (*) можно было бы упростить, отбросив первое слагаемое — оно практически не влияет на функцию отклика:

$$y=-10x_2.$$

Относительная ошибка при этом не превышает 0,02.

А теперь запишем подобное уравнение, но в общем виде

$$y=b_1x_1+b_2x_2,$$

где b_1, b_2 — оценки (а не точные значения) коэффициентов регрессии. Если b_1 и b_2 независимы, как случайные величины, то картина будет аналогичной: абсолютные величины $|b_1|$ и $|b_2|$ укажут быстроту изменения функции отклика при увеличении x_1 и x_2, а их знаки — направления этого изменения. Но если коэффициенты b_1, b_2 как-то связаны, то картина нарушается: погрешности в определении одного коэффициента могут повлиять на значения другого вплоть до изменения его знака. Теперь коэффициенты не служат уже мерой относительного влияния факторов, нельзя бездумно исключать из уравнения факторы с «маленькими» коэффициентами.

Впрочем, само уравнение регрессии не следует отправить на помойку: оно еще пригодится для предсказания значений функции отклика внутри исследуемой области.

Итак, у пассивного эксперимента много пороков. Конечно, если нет другого пути, как при изучении движения кометы, пользуются данными пассивного эксперимента. Но во многих задачах техники и естествознания исследователь имеет возможность отказаться от роли инертного наблюдателя и перейти от пассивного эксперимента к активному.

Активный против пассивного

Нелегко превратить дьявола в ангела добродетели и осудить пороки системы легче, чем указать способ их исправления или, тем более, построить «непорочную» систему.

Пороки пассивного эксперимента мы уже осудили, но исправить их нелегко. Все же можно многое сделать для повышения эффективности эксперимента, его «отдачи».

Давайте обсудим, чем же располагает экспериментатор, желающий проявить активность, что он может изменять, выбирать или отбрасывать и за что он должен бороться.

Прежде всего заметим: цель наблюдений, экспериментов, опытов — получение информации об изучаемом объекте, процессе, явлении. Следовательно, задача активного экспериментатора — добывание необходимой информации при наименьших затратах средств и времени или, если лимитированы средства и время, то его задача — получение наибольшей возможной информации при заданных ограничениях на средства и время.

Но какую информацию следует добыть? Явления природы бесконечно разнообразны, объекты техники сложны и, по-видимому, нет бесполезной информации, бесполезной вообще... Это как раз и есть философская база пассивности: что бы мы ни узнали, все на пользу. Однако активный экспериментатор — не зевака: он занят решением определенной задачи, и информация ему нужна не какая

попало, а именно та, которая даст возможность решить его задачу.

Итак, первое дело активного экспериментатора — это отбор информации, с одной стороны, нужной, необходимой для решения поставленной задачи и, с другой стороны, достаточной для ее решения. Если информации не хватает, ее надо добывать, а если нельзя добыть, то следует удовлетвориться возможным приближенным решением, но надо ясно себе представлять его ограниченность, неполноту. Излишняя информация не только бесполезна, но подчас и вредна: она не только занимает время и зря использует средства, но, бывает, создает ненужный фон, на котором может потеряться полезная информация, а то и создает предрассудки.

Таким образом, для отбора информации необходим четкий логический анализ задачи. При этом, говоря об информации, я имею в виду достаточно широкое ее понимание. Скажем, сама постановка задачи, ее ясная формулировка — это тоже информация о задаче.

Не подумайте, будто четкий логический анализ задачи — легкое дело, и не только логический анализ эксперимента, но и сама идея подобного анализа даются человечеству с трудом.

Каждая хозяйка хочет сварить хороший суп. Но чего же она должна добиваться? Ясно, суп не следует пересолить или недосолить, овощи должны быть проварены, но не разварены, пряности следует положить своевременно. А суп почему-то неудачный...

Если вам дадут попробовать борщ, приготовленный разными хозяйками, то сможете ли вы всегда сказать, какой из них лучше? Едва ли, даже если тому не будет препятствовать нежелание огорчить одну из женщин. И вместо однозначного ответа вы будете крутить известный словесный хоровод: «С одной стороны, нельзя не признать, с другой — нельзя не отметить...». Возникающие здесь трудности мы уже обсуждали в разделе «Немного о критериях», причем в вопросе о качестве борща, пожалуй, еще больше произвола, чем в проблеме выбора наилучшего из сочинений или более «высокого» из классов.

Когда изучается заводская установка, технологический процесс или сложный продукт химического производства, то приходится преодолевать те же трудности, нет, пожалуй, тут еще труднее выбрать правильную позицию. Какие из характеристик процесса или продукта взять в качестве

выходных переменных? Какие из измеряемых или управляемых переменных взять в качестве входов (в теории эксперимента их называют факторами), собрав все остальные в группу неконтролируемых переменных? Я уже упоминал о возникающих трудностях при изучении процесса первичной переработки нефти, где управляемых, измеряемых или контролируемых переменных порядка двухсот. С аналогичным положением мы встречаемся, к сожалению, весьма часто.

Лучше обсуждать конкретную проблему: она понятнее читателю, да и мне приятнее — паутина общих разговоров как-то обволакивает, и хочется забросить рукопись и заняться чем-нибудь содержательным.

Поэтому расскажу об одной реальной задаче, которой приходится заниматься последние годы мне и моим сотрудникам. На этой задаче я здесь остановился по двум причинам: во-первых, она мне хорошо знакома, и, следовательно, о ней можно рассказать лучше, чем о «чужой» задаче, и, во-вторых, на имеющемся материале удастся продемонстрировать многие из идей планирования активных экспериментов.

Все, конечно, знают пословицу: «не подмажешь — не поедешь». Хотя обычно используется ее переносный смысл, но и буквальный содержателен: для функционирования моторов, станков, различных агрегатов, где есть движущиеся части, необходима смазка. Даже немазаная телега не только вызывает раздражение неимоверным скрипом: из-за недостатка смазки ее скорость при тех же условиях падает.

В качестве одной из основных смазок используются технические масла, получаемые при переработке нефти.

При производстве технических масел на нефтяной основе широко применяется их очистка от нежелательных компонентов. Но даже с помощью самых эффективных методов очистки не удается получить масла с весьма высокими эксплуатационными характеристиками, которые требует сегодня промышленность. Масла должны быть устойчивыми к окислению, не вызывать коррозию металлов, поверхности которых они смазывают, трущиеся поверхности металлов не должны быстро изнашиваться, они должны обладать хорошими моющими свойствами. Для улучшения таких эксплуатационных свойств масел широко применяются специальные химические вещества — их называют присадками.

Добавление присадок в небольших количествах, от сотых долей процента до нескольких процентов, приводит к значительному повышению качества масел. Скажем, сотрудникам моей лаборатории совместно с технологами удалось выбрать соотношение присадок, повышающее срок службы масла одного типа более чем в два раза! Как понимаете, ради такого дела стоит поработать.

Казалось бы при современном состоянии химической науки должна существовать развитая теория, дающая возможность описать действие присадок, выбрать наилучшие присадки и их оптимальное процентное соотношение. Но такой теории пока нет. Поэтому и набор присадок, и их концентрации выбирают экспериментально.

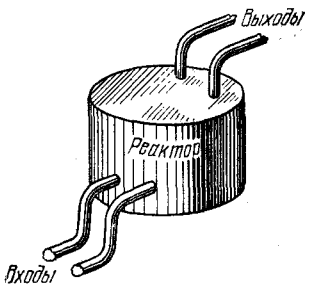


Рис. 26.

Я буду обсуждать сейчас только проблему выбора оптимального соотношения концентраций присадок. Схематически будем представлять себе всю ситуацию следующим образом. Имеется сосуд или, говоря более научно, реактор, частично заполненный основой — тем масляным полуфабрикатом, качественные показатели которого надо улучшить, оптимизировать. Сбоку в сосуд врезаны трубки, по каждой из них можно добавлять одну из присадок — это входы (рис. 26). Дозировка осуществляется с помощью открывания кранов на нужную величину. Естественно, в сосуде все тщательно перемешивается, и для определения качественных показателей жидкость поступает в трубки, указанные сверху,— это выходы. Будем считать, что на каждом из выходов измеряется свой качественный показатель. Такая схема — модель и любого другого технологического процесса: трубки — это входы, краны — управляющие устройства, изменяющие входы, а выходы — это верхние трубки.

В обсуждаемой задаче исследователь должен выбрать оптимальное соотношение концентраций присадок. Но что следует считать оптимальным соотношением? Технологи считают важными ряд показателей. Скажем, кислотное число характеризует стабильность масла к окислению кислородом воздуха, и надо постараться его сделать как можно меньше. Но и коррозию надо уменьшить; она характе-

ризуется удельной величиной коррозии, то есть относительным изменением веса металлов, подвергающихся коррозирующему воздействию жидкости. В технике используют сталь и чугун, медь и латунь, да и другие металлы и сплавы, а коррозионный эффект одной и той же жидкости на разные металлы может быть разным. На какую же ориентироваться?

Трение деталей приводит к их износу, который измеряется посредством условной единицы — диаметра пятна износа при движении шарика из изучаемого металла, конечно, дозируемого движения в специально созданном измерительном приборе.

Известно, что масло со временем густеет — увеличивается его вязкость, — и это существенно ухудшает эксплуатационные свойства масла. Следовательно, нужно добиваться, чтобы вязкость с течением времени не увеличивалась или росла как можно медленнее.

Многие приборы, аппараты, моторы имеют резиновые детали, соприкасающиеся с маслом, и это иногда приводит к набуханию резины. Набухание резины измеряют в процентах, и следует обеспечить возможно меньшую величину этого процента.

Можно указать еще ряд важных качественных показателей, но, думаю, вам уже ясна общая картина: выбранные показатели достаточно разнообразны.

Технологи экспериментально обнаружили важный факт: если добавлять сразу несколько присадок, то иногда наблюдается эффект усиления их действия,— они говорят, эффект синергизма, то есть содружественного действия в одном направлении.

За счет синергического эффекта можно надеяться, например, повысить стабильность масла к окислению либо значительно снизить расход присадок — они представляют собой сложные и весьма дорогостоящие химические вещества. Однако в других концентрациях присадки могут выступить как антагонисты. Выбранные качественные показатели могут быть как-то связаны между собой, причем, возможно, улучшение одного из них ведет к ухудшению другого. Следовательно, добиться оптимизации всех нужных показателей нельзя. Это не новая для нас ситуация: она уже подробно обсуждалась в разделе «Немного о критериях».

Обычно обстановка такова: потребители выдвигают какие-то требования, они подчас противоречивы, и их точ-

ное выполнение либо затруднительно, либо даже невозможно. И технологу-разработчику приходится идти на какие-то компромиссы.

В обсуждаемой задаче можно было бы выбрать какой-либо экономический критерий, например минимум приведенных затрат, но сейчас технологи поступают иначе. В качестве критерия выбирается характеристика качества масел, наиболее критичная в данной ситуации, например обладающая слишком значительным разбросом при производстве или выходящая за рамки допустимых технологических ограничений, или еще что-либо подобное.

Впервые мы столкнулись с задачей выбора оптимального соотношения концентраций присадок, когда перед технологами поставили задачу минимизации кислотного числа. Именно оно и было выбрано в качестве критерия качества. А на остальные выходные характеристики были заданы лишь технологические ограничения, то есть указывалось, в каких границах допустима вариация соответствующего параметра. Я не буду уточнять вид этих ограничений, моя задача — демонстрация методов, и ограничения — это просто некоторые неравенства на выходные параметры.

Теперь можно сформулировать задачу оптимизации: нужно определить соотношение концентраций присадок, обеспечивающее минимум кислотного числа, при выполнении заданных ограничений на остальные выходные переменные. Позже нам пришлось изменить критерий и оптимизировать другой выходной параметр, но существо задачи осталось тем же.

Конечно, прежде чем браться за такую задачу, исследователь изучает установку, технологический процесс или механизм явления. Да кроме того, он богат всякими знаниями. Какую часть из имеющейся в его распоряжении информации — априорной, доопытной информации — ему следует использовать? И этот вопрос совсем не прост: мы обладаем сведениями, ценность которых, их достоверность весьма неоднородны. Скажем, в проблеме выбора режима для бегуна на короткие дистанции тренер может опираться на свой опыт работы с бегуном на длинные дистанции или с метателем диска. Но ясно — не весь опыт может быть использован, коль решаются разные задачи. Но на какую часть имеющегося опыта можно опереться, а какая может принести вред, создавая различные заблуждения или предрассудки?

При решении задач естествознания или техники отбор априорной информации оказывается не менее трудным. Каким данным литературных источников можно доверять, а каким нет? Как воспользоваться данными, полученными при наблюдении аналогичного процесса, но на другой установке, при ином конструктивном решении, работающей на другом сырье, в других условиях, при использовании других веществ? А может быть, как раз следует ориентироваться на собственный опыт работы, хоть он и относится к совершенно непохожим объектам, но это ваш собственный опыт, многократно продуманный, пережитый. Ведь истинный талант — это как раз возможность синтезировать идеи на основании аналогий из разных областей, из двух или трех...

Я не могу дать рецепт поведения исследователя при отборе априорной информации — это не формализуемый процесс, как само научное творчество, а помочь здесь может лишь здравый смысл и опыт научной работы.

Но вот мы уже, пользуясь отобранной априорной информацией, выбрали входы — факторы и выходные переменные, выбрали критерий качества и должны проводить эксперименты. Сколько опытов надо провести? При каких значениях факторов, то есть в каких точках допустимой области факторного пространства следует ставить эксперименты? В какой последовательности нужно брать точки для проведения экспериментов?

Обсудим сначала вкратце первый из вопросов — оценим «размеры бедствия». Простейшее рассуждение таково: надо тщательно обследовать область возможного изменения факторов, а для этого нужно изменять по отдельности каждый из факторов достаточно часто в интервале его изменения, закрепляя другие, и постепенно перебрать все возможности. Конечно, действуя так, мы найдем оптимум. Но прикинем, сколько же при этом экспериментов нужно провести.

Основным здесь является число различимых значений. Если фактор x изменяется в интервале от 0,1 до 3,3, а измеряется он с точностью до 0,2, то различимых значений у него будет

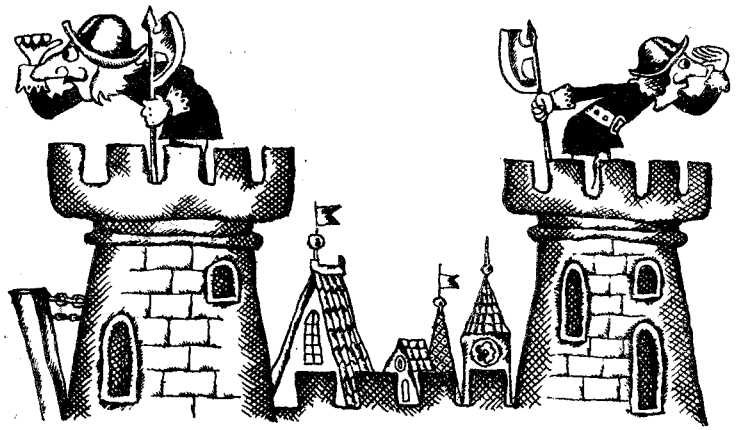
$$\frac{3,3-0,1}{0,2}=16.$$

Ясно, что более подробные измерения просто не нужны. Варьируемые в опыте значения факторов называют в теории эксперимента *уровнями*. Выбор числа уровней для всех факторов определяет число различимых значений входов. Если при подборе оптимальных концентраций присадок используется пять присадок, и каждая варьируется по пяти уровням, то общее количество различимых состояний будет равно $5^5=3125$.

При экспериментальном определении изучаемых характеристик масел опыт может длиться две недели, и для обследования 3125 точек факторного пространства понадобится более 12 лет, в течение которых не только часть исполнителей уйдет на пенсию, но и изменятся требования к маслам — естественное следствие быстрого развития техники. Итак, перебор всех различимых значений нас не устраивает. Что же делать? Не отказываться же от возможности заметно улучшить качественные показатели масел из-за невозможности осуществить полный перебор вариантов?

Вот и создается ситуация, кажущаяся трудно преодолимой: или проводить весьма большое число опытов, затрачивая на работу месяцы и годы, или провести немного опытов, выбрать среди полученных результатов лучший и закрыть глаза на реальную возможность найти комбинацию факторов, в которой показатель качества будет значительно превосходить полученный, добытый случайно, ибо довольно случайно была выбрана точка. Для оправдания всегда можно многозначительно сказать: «Нельзя объять необъятное, а мы мол все же нашли комбинацию факторов лучшую, чем наши предшественники».

Но при такой стратегии эксперимента — в радиотехнике ее презрительно называют «методы тыка» — невелики шансы найти, нет, не найти, а случайно наткнуться на оптимум. Вот здесь на помощь экспериментатору и приходит логический анализ эксперимента! И в ближайших разделах я подвергну анализу возможности экспериментатора и укажу стратегии эксперимента, не только значительно более выигрышные, чем «методы тыка», но и приводящие к успеху.



Откуда лучше видно?

В условиях активного эксперимента исследователь имеет возможность в факторном пространстве сам выбирать точки, в которых будут ставиться эксперименты. Это, конечно, огромное преимущество по сравнению с пассивным экспериментом, где приходится довольствоваться лишь теми точками, «что бог послал», и именно из-за этого «божьего» произвола проистекают пороки пассивного эксперимента.

Однако разумный, целесообразный выбор точек в факторном пространстве — задача не из простых, и в большинстве работ по теории планирования экспериментов обсуждается проблема выбора точек в факторном пространстве.

Отобранные точки в факторном пространстве и последовательность, в которой ставятся эксперименты в выбранных точках, называют планом эксперимента, и поэтому выбор точек и стратегии их последовательного использования — это и есть планирование эксперимента.

Начну обсуждение планирования эксперимента с практической задачи, решением которой несколько лет назад совместно с инженерами-гидростроителями занимался мой бывший аспирант Я. А. Коган, да и я тоже принимал некоторое участие. Она относится к вопросу автоматической корректировки состава бетонной смеси и имеет особенно

большое значение при строительстве высоких бетонных плотин.

Бетон должен быть однороден. Цемент, конечно, дороже песка или гравия, и количество цемента нужно по возможности уменьшить, впрочем, не только из экономических соображений, но и для снижения экзотермических реакций. Объемы бетонных работ при строительстве гидроэлектростанций велики. Все это вместе диктует необходимость автоматической корректировки состава бетонной смеси и оправдывает применение относительно сложного и дорогого оборудования автоматического управления.

Разработка системы корректировки состава бетонной смеси опирается на два положения. Во-первых, изменение характеристик бетона и бетонной смеси решающим образом зависит от влажности заполнителей, главным образом песка. Во-вторых, не менее важную роль играют свойства набора заполнителей (песок, гравий, щебень, шлак) — гранулометрического состава: в конечном счете они определяют водопотребность бетонной смеси данного состава — водосодержание при заданной консистенции.

За рубежом были созданы системы, корректирующие дозу воды в зависимости от влажности песка. Такие системы могут быть эффективны при сохранении высокой однородности свойств заполнителей или относительно небольшой скорости их изменения и при колебаниях влажности крупного заполнителя в небольших пределах.

В отечественной практике сохранить высокую однородность свойств заполнителей более или менее возможно лишь на бетонных заводах небольшой производительности. Но в условиях нашего крупного, а то и гигантского гидростроительства, при весьма высоком темпе ведения бетонных работ, на заводах большой производительности обеспечить такую однородность свойств заполнителей практически невозможно.

Например, наблюдения, проведенные на строительстве Красноярской ГЭС, свидетельствуют о большой скорости изменения влажности и гранулометрического состава заполнителей, особенно песка. В этих условиях целесообразно применять системы корректировки, учитывающие колебания влажности и гранулометрического состава заполнителей.

Задача, таким образом, состояла в создании активной системы, автоматически определяющей оптимальное по водопотребности соотношение компонентов бетонной смеси.

Очевидно, что каждому набору заполнителей (гранулометрическому составу) будет соответствовать определенная минимально возможная водопотребность бетонной смеси. При этом основное влияние на водопотребность оказывает содержание песка в смеси заполнителей.

Инженеры-гидростроители знают по опыту зависимость между истинной водопотребностью и долей песка в смеси заполнителей — она близка к параболической. Экстремум параболы соответствует минимально возможной водопотребности. Если обозначить через v водопотребность и через x — долю песка, то, как показывают наблюдения, зависимость между ними задается формулой

$$v=b(x+a)^2+c, \qquad (*)$$

где $b>0$, c и a — неизвестные постоянные, которыми определяются форма и положение параболы. При изменении гранулометрического состава заполнителей положение и форма параболы изменяются и точка a — координата ее экстремума — сдвигается.

Таким образом, задача автоматического управления сводится к поддержанию консистенции бетонной смеси в заданном интервале и нахождению той оптимальной доли песка, при которой водопотребность минимальна. Допустимая доля песка x изменяется в некоторых пределах $x_{\min}\leqslant x\leqslant x_{\max}$. Для определения неизвестного нам значения параметра a в формуле (*) можно поступить очень просто: задав три значения x_1, x_2, x_3 переменной x внутри интервала $(x_{\min}, x_{\max})$, следует определить экспериментально соответствующие значения v_1, v_2, v_3 функции v и затем найти из системы уравнений

$$\left.\begin{aligned} v_1 &= b(x_1+a)^2+c,\\ v_2 &= b(x_2+a)^2+c,\\ v_3 &= b(x_3+a)^2+c \end{aligned}\right\} \qquad (**)$$

значения параметров a, b и c. Решение этих уравнений относительно a — школьная задачка, и совсем не здесь нас ждут *осложнения*. А *они* есть, и в *общем-то из-за них* я и повел весь этот разговор.

При решении системы уравнений (**) все равно, какие значения x_1, x_2, x_3 выбраны внутри допустимого интервала $(x_{\min}, x_{\max})$. Но при экспериментальном определении значений v_1, v_2, v_3 всегда имеется погрешность, ошибка, и ее учет резко меняет всю картину.

Давайте еще немного упростим задачу, положив $b=1$, $c=0$, так что уравнение примет вид

$$v=(x+a)^2, \qquad (***)$$

и пусть независимая переменная x изменяется в пределах от $x_{\min}=-1$ до $x_{\max}=+1$; этого всегда можно добиться простой заменой переменных.

Итак, требуется *найти положение экстремума параболы* (***), когда x изменяется в пределах $-1\leqslant x\leqslant+1$. Теперь для этого достаточно определить два значения v_1, v_2 при некоторых x_1, x_2, и ясно, что значения x можно выбрать произвольно на интервале $(-1, +1)$. Но какие значения x_i выбрать, если измерения v производятся с ошибкой? Мало того: теперь возникает вопрос о необходимом количестве точек x_i, в которых производятся измерения, и количестве измерений в каждой из них. Здесь уже надо переформулировать задачу.

Будем предполагать, что значения $v_i=(x_i+a)^2$ измеряются со случайной аддитивной ошибкой ε_i, то есть в результате измерений получаются величины

$$y_i = v_i + \varepsilon_i.$$

Ошибки в определении v_i влекут за собой ошибки в определении интересующей нас величины a, и естественно поставить перед собой задачу выбора плана эксперимента, при котором величина a будет определена с наименьшей возможной ошибкой. Здесь, так же как в задаче выбора оптимальной стратегии безбилетником, о которой мы говорили в разделе «Риск», желательно исключить зависимость ошибки от параметра a. Для этого вводится его априорное распределение, и ошибка еще раз усредняется по априорному распределению.

Я сознательно не записал формулы: получается весьма сложная задача вариационного исчисления. Но попробуем подобраться к оценке параметра a с более простых и наглядных позиций.

Естественно полагать величины ошибок ε_i независимыми как от значений x_i, так и в различные моменты времени, то есть при различных значениях i. Как же выбрать значения независимой переменной x_i, обеспечивающие высокую точность определения параметра a? При независимости величины ошибок ε_i от x_i относительное значение величины ошибки зависит от величины v_i, и теперь

ясно, что при разных x_i относительная величина ошибки будет различна. Следовательно, вполне разумно поставить задачу выбора тех значений x в допустимом интервале, при которых относительная величина ошибки была бы минимальной.

Местоположение точки a на интервале $(-1, +1)$ неизвестно, и парабола «общего положения» выглядит так, как представлено на рис. 27. Наибольшего значения функция $v(x)=(x+a)^2$ достигает, очевидно, на одном из концов интервала $(-1, 1)$, то есть либо в точке $+1$, либо в точке -1, но вот, в какой из этих точек достигается наибольшее значение, нам неизвестно. Поэтому нужно использовать оба крайних значения.

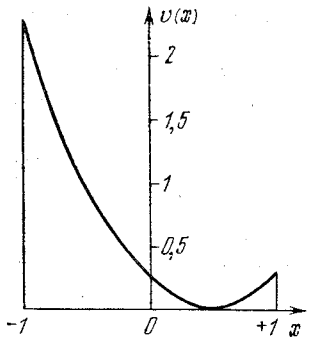


Рис. 27.

Если сделать только два измерения — по одному в точках $x=-1$ и $x=+1$, то получим

$$y_1=(-1+a)^2,\ y_2=(1+a)^2,$$

откуда после элементарных преобразований получим $a=(y_1-y_2)/4$.

Однако можно проводить измерения многократно. Та же логика приводит к выводу о целесообразности проводить измерения, по-видимому, лишь в крайних точках -1 и $+1$, ибо измерения во внутренних точках интервала $(-1, +1)$ могут дать значения относительных величин ошибок, превосходящие ту из ошибок, которая относится к наибольшему из крайних значений $v(-1)$ или $v(+1)$. Я написал «по-видимому» не просто так: наше рассуждение довольно правдоподобно, но не доказано. На самом же деле оно неверно: оптимальная стратегия оказывается иной. Однако стратегия измерений, при которой половина измерений производится в точке -1, а вторая половина в точке $+1$ — будем ее называть субоптимальной стратегией,— оказывается достаточно хорошей. Я уже говорил об оптимальной стратегии: определение значений x_i, в которых надо производить измерения, приводит к необходимости решения трудной задачи вариационного исчисления. Сравнение субоптимальной стратегии с оптимальной

показывает, что применение субоптимальной стратегии при $2n$ измерениях (по n измерений в каждой из крайних точек интервала $(-1, +1)$) дает значение среднеквадратической погрешности меньшее, чем при n измерениях и использовании оптимальной стратегии. Этот факт подтверждает сразу два интуитивно понятных положения: во-первых, важность и трудность выбора тех точек x_i, в которых следует производить измерения, и, во-вторых, значимость в обсужденном примере с параболой крайних точек -1 и $+1$.

Именно на этой субоптимальной стратегии и была построена система автоматической корректировки бетонной смеси.

Итак, при обдумывании эксперимента, при его планировании нужно тщательно взвесить, при каких значениях независимой переменной его следует производить. Конечно, разобранный пример не дает возможности предугадать разнообразие ситуаций, грозящих экспериментатору, и поэтому к вопросу о том, где ставить эксперименты, я еще вернусь.

Шаг за шагом

Давайте вспомним о туризме. Вот группа туристов собирается взобраться на вершину горы. Туристы, конечно, выражаются более высокопарно — они говорят «осуществить восхождение», после чего прогулка на вершину горы из отдыха превращается в спорт с достижениями различной степени трудности.

Вы, удобно расположившись в кресле недалеко от горы, следите в бинокль за действиями туристов. Группа остановилась и с помощью какого-то современного прибора определила свою высоту над уровнем моря. Затем они собрали прибор, взгромоздили на спины традиционные рюкзаки и сделали пятьсот шагов строго на восток. Вновь вынули прибор, измерили высоту над уровнем моря, собрали вещи, надели рюкзаки и сделали пятьсот шагов, но уже строго на север, на запад и на юг. Перейдя в одно из отмеченных мест, они повторяют процедуру, постепенно определяя высоту над уровнем моря в различных точках сетки с расстоянием между точками в пятьсот шагов.

Вечером, когда группа, усталая, но весьма удовлетворенная своими успехами, возвращается на турбазу, вы пытаетесь понять, зачем же они ходили туда-сюда и измеряли высоты. Туристы снисходительно вам объясняют: задача группы — совершить восхождение на вершину горы. Вершиной называется самая высокая точка, и, е с т е с т в е н н о, нужно измерить высоту достаточно большого количества точек на поверхности горы и выбрать среди них самую высокую. Она и будет принята за вершину.

Вы, внимательно слушая объяснения, смотрите в лучистые глаза туристов и думаете о необходимости вызова психиатра: по-видимому, наблюдается случай коллективного психоза...

К счастью, туристы на самом деле действуют иначе, но экспериментаторы весьма часто поступают именно описанным методом, и никто не подозревает нарушений в их психике.

Почему же происходит такая бессмысленная трата времени и средств?

Выше я выделил разрядкой слово «естественно»: хотелось подчеркнуть ход рассуждений... Да, можно определить самую высокую точку на поверхности, если обследовать систематически высоту точек, лежащих на поверхности, и выбрать самую высокую. Но рассуждение это далеко не единственно возможное и совсем не вытекающее естественно из постановки задачи как само собой разумеющееся. Как раз наоборот — это плохой и неестественный ход рассуждений.

Туристы обычно поступают так: поглядев на гору, они выбирают удобный путь на вершину и вовсе не обследуют все точки, принадлежащие какой-либо сетке.

Я думаю, что экспериментаторы подчас устраивают перебор точек, будь то на сетке или по какому-либо другому правилу, вследствие отсутствия привычки к четкой формулировке задачи. Одно из достижений в экспериментальных науках, полученное благодаря проникновению методов математической статистики в эксперимент,— это принятие на вооружение экспериментатора методов четкого логического анализа процесса выдвижения гипотез, постановок задач и принятия или отвержения гипотез.

Вернемся теперь к задаче подбора оптимального соотношения концентраций присадок к маслам. Напомню важный тезис: успех статистического анализа определяется разумностью постановки задачи.

В обсуждаемой задаче надо найти оптимальную точку — такое соотношение концентраций, которое дает наилучший эффект. Зачем же нужно обследовать всевозможные концентрации или хотя бы весьма большое их количество? Логика говорит о другом: нужно указать правило проведения экспериментов, которое дает возможность за минимальное число опытов найти искомое оптимальное

соотношение концентраций. При такой постановке уже совсем глупо обследовать подробно область всех возможных концентраций, столь же глупо, как измерять высоты большого числа точек на склонах горы вместо выбора пути для подъема на ее вершину.

Быть может, и не самый удобный, но самый короткий путь к вершине — это в каждой точке двигаться по направлению, наиболее круто идущему вверх. Если вершина одна, по дороге нет других маленьких, как говорят, локальных вершин, то такой путь приведет наверняка к успеху, из какой бы точки на склоне вы ни начали подъем.

Все же есть разница между туристом и технологом: технолог не видит свою гору — поверхность отклика, и это создает внешнее впечатление, будто он находится совершенно в другой обстановке. Однако это ложное впечатление.

В самом деле, давайте поставим туриста и технолога в сходные условия, для чего предложим туристу подняться на вершину безлунной ночью. Стратегия его поведения довольно очевидна: он будет совершать небольшие шаги вправо — влево, вперед — назад и переходить туда, где подъем наиболее ощутим. Так он и дойдет до вершины Определить факт достижения вершины тоже нетрудно: если все четыре шага вправо и влево, вперед и назад приводят к спуску, то, значит, турист достиг вершины. Конечно же не может быть и речи о проверке высоты всех точек на сетке и выборе наивысшей — это же, очевидно, ничем не оправданный, а потому и излишний труд.

Точно так же следует поступать и технологу при поиске оптимального соотношения концентраций присадок. Он должен прежде всего заняться планированием эксперимента.

Напомню: в качестве критерия оптимальности композиции присадок было выбрано общее кислотное число, и оптимальной композицией считалась та, при которой общее кислотное число достигало минимума. Таким образом, за независимые переменные — факторы — принимаются сами концентрации присадок, а выход или функция отклика — общее кислотное число.

При такой формализации задача состоит не в поиске максимума (вершины на поверхности), а в определении минимума (самой низкой точки). Логически поиск минимума или максимума осуществляется совершенно одинаково: муравью все равно, лежит ли шляпа полями вниз или

вверх: его кратчайший путь от бантика на канте до вершины — не случайно же вершину шляпы называют донышком — один и тот же (рис. 28).

Когда движение все время происходит по пути, наиболее круто идущему вниз, есть гарантия либо достигнуть минимума, либо выйти на границу допустимой области. Но здесь возможны весьма серьезные осложнения.

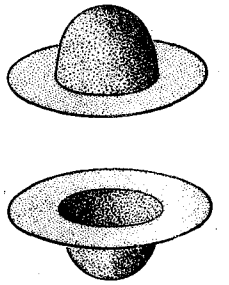


Рис. 28.

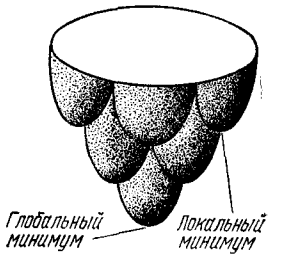


Рис. 29.

Если не ввести никаких предположений о структуре поверхности отклика, то положение наше будет безнадежным.

Для наглядности пусть у нас всего два фактора, так что поверхность отклика — это обычная поверхность. Скажем, если не предполагать непрерывность поверхности, то она может вести себя совершенно нерегулярно, и никаких максимумов и минимумов в понятном наглядном смысле у нее не будет. Но пусть поверхность достаточно гладкая, без отвесных стен или обрывов. Однако она может быть похожей издали на гроздь винограда (рис. 29): у нее много минимумов — математики их называют локальными,— и, попав в один из них, скажем, отмеченный жирной точкой, нелегко догадаться о наличии других минимумов, еще более «низких». И тем более трудно достигнуть самой низкой точки — глобального минимума.

Не хочется верить, будто природа столь коварна. Более естественно, хотя бы сначала, предполагать поверхность отклика устроенной как-то более просто, без подобной многоэкстремальности, хотя, конечно, нельзя априори исключить вообще возможность нескольких локальных экстремумов.

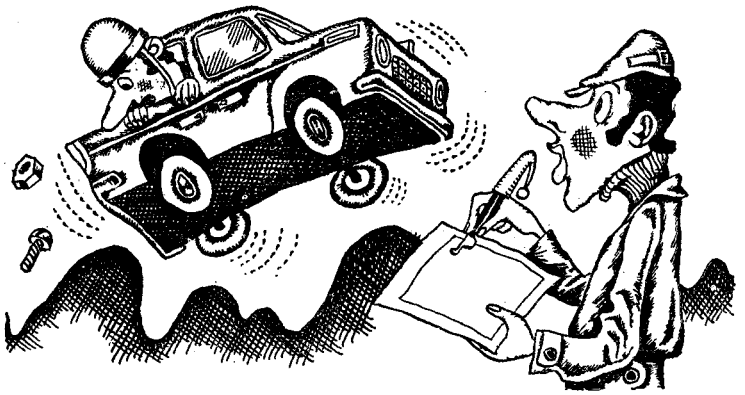
Выделим на поверхности какую-то точку и рассмотрим небольшую ее окрестность. В малой окрестности кусок

достаточно гладкой поверхности практически не отличим от куска плоскости, и если плоскость не параллельна горизонтальной плоскости, то по ней можно сделать шаг вниз по наиболее крутому направлению. Перейдя в новую точку и выделив ее небольшую окрестность, можно опять построить кусок плоскости, практически неотличимый от куска поверхности, и сделать вновь наиболее крутой шаг вниз. Так будет продолжаться движение вниз до тех пор, пока либо не будет достигнута граница допустимой области, либо в окрестности точки плоскость окажется параллельной горизонтальной плоскости. Такая точка, как вы, вероятно, помните, называется стационарной. Латинское слово «стационарный» буквально означает «стоящий на часах», а в переносном смысле — неподвижный, постоянный, остающийся на месте. В математическом анализе стационарность понимается в последнем смысле: если в стационарную точку на поверхности поставить тяжелый шарик, то он останется на месте, а из любой нестационарной точки шарик покатится по поверхности. Если стационарная точка — это минимум, то после любого маленького отклонения шарик возвратится на прежнее место. Однако стационарными будут не только точки минимума или максимума, но и точки, в окрестности которых поверхность имеет седлообразный вид или оказывается цилиндрической с направляющей, параллельной горизонтальной плоскости.

Для выяснения поведения поверхности в окрестности стационарной, но «подозрительной» точки ее следует заменить уже поверхностью второго порядка, определить ее вид и, наконец, вздохнуть с облегчением, если обнаружится минимум. Впрочем, и этот вздох может оказаться преждевременным: мы ищем минимум глобальный, и следует проверить, не попали ли мы в локальный минимум.

Как видите, предлагаемая шаговая или последовательная стратегия эксперимента представляет собой вариант последовательного анализа Вальда, где для проверки гипотезы (о стационарности точки, о наличии минимума, о глобальности этого минимума) производятся эксперименты. Я не буду более подробно останавливаться на методах поиска экстремума — они, например, изложены популярно и изящно в прекрасной книжке А. А. Первозванского *).

*) А. А. Первозванский, Поиск, «Наука», 1970.



Где ставить эксперименты?

Этот вопрос мы обсудили в разделе «Откуда лучше видно?» лишь для ситуации с одним фактором при параболической функции отклика. А теперь обратимся к многофакторным задачам.

Начну с простой задачи взвешивания трех предметов или образцов, безразлично каких *). Назовем образцы A, B, C и их веса так же — путаницы не будет. Первое, что приходит на ум, это последовательно взвешивать каждый из образцов. Так и поступает традиционный исследователь, но вначале он делает холостое взвешивание для определения нулевой точки весов. Когда образец кладется на весы, в таблице ставится $+1$, когда он на весах отсутствует — ставится -1. Результат взвешивания обозначим через y с соответствующим индексом (см. таблицу 4).

Итак, здесь исследователь изучает поведение каждого фактора в отдельности, то есть проводит однофакторные эксперименты. Вес каждого из образцов оценивается лишь по результатам двух опытов: холостого опыта и того, в котором изучаемый объект был на весах. Например, вес

*) В. В. Налимов, Теория эксперимента, «Наука», 1971. Из этой острой и интересной книги, правда обращенной к достаточно подготовленному читателю, я многое почерпнул и здесь использовал, в том числе и задачу о взвешивании.

образца A равен

$$A=y_1-y_0.$$

Как обычно, ошибка взвешивания предполагается независимой от взвешиваемой величины, аддитивной и

Таблица 4

Традиционная схема взвешивания трех образцов

Номер опыта	A	B	C	Результат взвешивания
1	−1	−1	−1	y_0
2	+1	−1	−1	y_1
3	−1	+1	−1	y_2
4	−1	−1	+1	y_3

имеющей одно и то же распределение. Тогда дисперсия измерения веса образца равна

$$D(A)=D(y_1-y_0)=Dy_1+Dy_0=2\sigma^2,$$

где σ^2 — дисперсия любого взвешивания.

Такой же будет дисперсия веса образцов B и C.

Но эксперимент можно провести и по другому плану — многофакторному: он представлен в таблице 5.

Таблица 5

Многофакторный план взвешивания трех образцов

Номер опыта	A	B	C	Результат взвешивания
1	+1	−1	−1	y_1
2	−1	+1	−1	y_2
3	−1	−1	+1	y_3
4	+1	+1	+1	y_4

Теперь нет холостого взвешивания. В первых трех опытах последовательно взвешивают образцы A, B, C, а в четвертом взвешиваются все три образца вместе.

Умножая элементы последнего столбца таблицы последовательно на элементы столбцов A, B, C и деля на два ибо, в соответствии с планом, каждый из образцов взвешивается дважды, получим веса

$$A=\frac{1}{2}(y_1-y_2-y_3+y_4),$$

$$B=\frac{1}{2}(-y_1+y_2-y_3+y_4),$$

$$C=\frac{1}{2}(-y_1-y_2+y_3+y_4).$$

Здесь веса образцов не искажаются весами других, ибо, например, в выражение для веса образца B каждый из весов образцов A и C входит дважды и притом с разными знаками. Дисперсия ошибки взвешивания теперь равна

$$D(A)=D\left(\frac{y_1-y_2-y_3+y_4}{2}\right)=\frac{1}{4}\cdot 4\sigma^2=\sigma^2,$$

то есть вдвое меньше, чем при однофакторном плане взвешивания. Если бы мы захотели при однофакторном плане получить такую же дисперсию, как и при обсуждаемом многофакторном, то следовало бы провести каждый из четырех однофакторных опытов дважды, то есть провести восемь взвешиваний вместо четырех.

Итак, при многофакторном плане каждый вес вычисляется по результатам всех четырех опытов — вот причина уменьшения дисперсии вдвое!

Не подумайте, будто мы потратили зря время на обсуждение такой тривиальной задачи. Точно то же самое будет при изучении функции отклика, линейно зависящей от трех факторов x_1, x_2, x_3.

Вспомним процесс электрообессоливания, где количество солей на выходе электрообессоливающей установки (y) зависит от количества солей на входе (x_1), количества добавляемого деэмульгатора (x_2) и времени пребывания эмульсии в электрическом поле (x_3). Когда эти факторы изменяются в некоторых определенных пределах, зависимость y от x_1, x_2, x_3 будет линейной.

Уравнение регрессии здесь имеет вид

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3. \qquad (*)$$

Будем варьировать каждый из факторов на двух уровнях, принимая за уровни наибольшее и наименьшее значения фактора в интервале его изменения и приписывая этим уровням кодовые обозначения, соответственно +1 и −1. Впрочем, как было указано в разделе «Откуда лучше видно?», с помощью линейной замены переменных можно добиться того, чтобы фактор имел интервал изменения (−1, +1).

Теперь для постановки экспериментов можно воспользоваться матрицей планирования, приведенной в таблице 5. Перепишу ее в новых переменных.

Здесь добавлен столбец фиктивной переменной x_0, нужной для оценки свободного члена β_0.

В соответствии с планом, эксперименты проводятся в следующих точках факторного пространства: в первом опыте фактор x_1 находится на верхнем уровне, а x_2 и x_3 на нижнем, то есть в преобразованных переменных эксперимент проводится в точке $(+1, -1, -1)$; во втором опыте x_2 на верхнем уровне, а x_1 и x_3 на нижнем, то есть в точке $(-1, +1, -1)$, и аналогично в третьем опыте — в точке $(-1, -1, +1)$, и в четвертом опыте — в точке $(+1, +1, +1)$.

После реализации плана получаются четыре уравнения с четырьмя неизвестными. Их решения и дадут оценки всех четырех коэффициентов регрессии β_0, β_1, β_2, β_3. Итак, в плане таблицы 6 число опытов равно числу определяемых констант. Такие планы называют насыщенными.

Таблица 6

Матрица планирования для линейной модели с тремя независимыми переменными

Номер опыта	План				Результаты эксперимента
	x_0	x_1	x_2	x_3	
1	+1	+1	−1	−1	y_1
2	+1	−1	+1	−1	y_2
3	+1	−1	−1	+1	y_3
4	+1	+1	+1	+1	y_4

Но заметьте: мы же использовали не все точки с «крайними» координатами, то есть с координатами ± 1, или, говоря другими словами, не все возможные комбинации выбранных уровней. В самом деле, всех возможных комбинаций из трех символов, каждый из которых принимает значения либо $+1$, либо -1, будет, как легко проверить, $2^3=8$. Мы пока использовали лишь 4 из них. А что же остальные?

Для ответа на этот вопрос давайте обратимся к еще более простой ситуации, когда факторов всего два, и варьируются они тоже на двух уровнях. План, задаваемый

всеми возможными комбинациями двух уровней,— он называется полным факторным, будет содержать $2^2=4$ точки, и они представлены в таблице 7 двумя средними столбцами.

Таблица 7

Матрица планирования для полного факторного эксперимента типа 2^2

План			
x_0	x_1	x_2	$x_1 x_2$
+1	−1	−1	+1
+1	+1	−1	−1
+1	−1	+1	−1
+1	+1	+1	+1

Если провести опыты согласно такому полному факторному эксперименту, то можно оценить все коэффициенты в уравнении регрессии

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_{12} x_1 x_2.$$

Последнее слагаемое здесь уже нелинейно, оно содержит произведение факторов, и поэтому его называют эффектом взаимодействия, хотя взаимодействие может быть и значительно более сложным. Но такова принятая сейчас терминология. Таким образом, полный факторный эксперимент дает возможность оценить коэффициенты более общего уравнения, чем линейное с двумя переменными.

Когда есть серьзные основания предполагать, что $\beta_{12}=0$, то в матрице таблице 7 можно приравнять $x_1x_2=x_3$ и тогда получится матрица таблицы 6, то есть план для трехмерного пространства факторов, но уже не полный факторный план для трех переменных, а его часть. Такой эксперимент называют дробным факторным экспериментом, а его план — дробной репликой, так что таблица 6 — это дробная реплика полного факторного плана типа 2^3, матрица которого представляется в виде второго, третьего и четвертого столбцов таблицы 8.

Теперь заметим: матрица планирования для линейной модели с тремя переменными, представленная в таблице 6, есть часть матрицы последнего плана — она состоит из четырех строк первых четырех столбцов и обведена

пунктиром. Поэтому план таблицы 6 называется полурепликой полного факторного эксперимента и обозначается 2^{3-1}. Если у факторов этой полуреплики поменять все знаки на обратные, то получатся нижние четыре строчки той же матрицы — это будет другая полуреплика.

Таблица 8

Матрица планирования для полного факторного плана типа 2^3

№ эксперимента	x_0	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
1	+1	+1	−1	−1	−1	−1	+1	+1
2	+1	−1	+1	−1	−1	+1	−1	+1
3	+1	−1	−1	+1	+1	−1	−1	+1
4	+1	+1	+1	+1	+1	+1	+1	+1
5	+1	−1	+1	+1	−1	−1	+1	−1
6	+1	+1	−1	+1	−1	+1	−1	−1
7	+1	+1	+1	−1	+1	−1	−1	−1
8	+1	−1	−1	−1	+1	+1	+1	−1

Теория планирования эксперимента началась с работ знаменитого английского статистика сэра Рональда Фишера в 20—30-х годах XX века и затем, в рассматриваемом направлении, была развита в пятидесятых годах в США Дж. Боксом и его сотрудниками, и именно эти последние работы, имевшие четко выраженный прикладной характер, стимулировали их широкое признание. Но терминология, принятая Боксом, не представляется очень удачной, ибо многие известные понятия, для которых уже есть устоявшаяся терминология в теории управления или в математической статистике, здесь названы по-иному. Принятая у нас терминология в теории планирования эксперимента — это либо переводы терминов с английского, либо просто их перенос, как, например, ротатабельный план, то есть план эксперимента, который обладает одинаковой погрешностью предсказания в любом направлении.

В обычном русском языке «реплика» означает замечание на слова собеседника или выражение. Такое содержание едва ли соответствует введенному термину. Но, скажем, в пьесах Бернарда Шоу роль часто состоит из реплик — их бросает актер по ходу действия, и в этом понимании реплика — часть роли, так же, как дробная реплика — часть полного факторного плана.

В английском языке слово «replicate» используется в живописи, где означает «повторять, копировать». Возможно, что авторы термина имели в виду этот смысл, вводя понятие дробных реплик полного факторного эксперимента. В русском контексте термин «реплика» мне не кажется удачным, хотя я, как видите, и попробовал найти ему оправдание.

Впрочем, хороша ли или неудачна терминология, у тех, кто пользуется сегодня теорией планирования эксперимента, нет другого выхода, как пользоваться ею, ворча и стеная, или приняв ее снисходительно, в зависимости от характера.

Полный факторный эксперимент типа 2^3 (таблица 8) дает возможность оценить коэффициенты уравнения регрессии, содержащего уже три парных взаимодействия и одно тройное. Соответствующие произведения записаны в верхней строке таблицы 8, и я не буду выписывать длинное уравнение регрессии — надеюсь, вам понятен его вид. Если же у исследователя есть уверенность в линейности поверхности отклика, то есть в отсутствии нелинейных членов в уравнении регрессии, то он вводит новые переменные $x_4=x_1x_2$, $x_5=x_1x_3$, $x_6=x_2x_3$, $x_7=x_1x_2x_3$ и получает матрицу планирования для оценки восьми коэффициентов (включая β_0) в линейном уравнении регрессии с семью факторами.

Если при решении задачи можно ограничиться линейным приближением, то в полном факторном эксперименте оказывается много «лишних» опытов. Скажем, при трех факторах, как мы уже видели, для вычисления оценок коэффициентов регрессии в линейном уравнении достаточно четырех опытов, а в полном факторном эксперименте типа 2^3 их восемь, и, следовательно, есть четыре «лишних». Результаты этих опытов могут быть использованы двояко: во-первых, с их помощью можно получить более точные оценки коэффициентов регрессии, во-вторых, можно проверить адекватность построенной модели. Но при семи факторах полный факторный эксперимент на двух уровнях содержит $2^7=128$ опытов, а, как я только что говорил, для оценки восьми коэффициентов линейного уравнения регрессии нужно всего восемь опытов. Таким образом, остается 120 «лишних», и, конечно, нет необходимости их все реализовывать. Достаточно лишь несколько из них использовать для проверки адекватности и уточнения оценок.

Подобные рассуждения можно продолжать и дальше, но, думаю, вам общая схема понятна, а детали... На самом деле здесь есть огромное разнообразие планов, различные подходы и возможности резкого сокращения и количества необходимых экспериментов и получения более полной и надежной информации. Но эта книжка — не учебник по теории планирования эксперимента, и я за дальнейшими подробностями отсылаю вас к специальной литературе.

Важно лишь отметить большие преимущества уже приведенных планов. Например, все коэффициенты в уравнении регрессии оцениваются независимо друг от друга. Это означает, что коэффициенты, скажем, в уравнении (*) указывают относительную величину соответствующих слагаемых, и, следовательно, можно пренебречь слагаемыми, маленькими по сравнению с другими. Иначе говоря, факторы с относительно маленькими коэффициентами можно отбросить как незначимые, не пересчитывая после этого заново коэффициенты.

Если поверхность отклика в изучаемой окрестности существенно нелинейна, то обойтись планированием лишь на двух уровнях не удается, и приходится использовать три уровня, да и минимальное количество экспериментов тоже приходится увеличить. Я не буду вдаваться в общие рассуждения, а вновь вернусь к проблеме выбора оптимального соотношения концентраций присадок к маслам.

Предварительный анализ показал, что критерий качества — общее кислотное число — нелинейно зависит от концентраций присадок, и, следовательно, поверхность отклика может быть сложно устроена.

В то же время лишь две из пяти присадок заметно изменяли кислотное число при вариации концентраций, а вариации остальных почти не влияли на выбранный критерий. Поэтому задача свелась к двухфакторной, и я могу продемонстрировать наглядно последовательные ситуации при поиске оптимальных концентраций присадок. Не буду обременять вас сложными названиями или принятыми обозначениями присадок — пусть они будут D и E.

В качестве первого шага исследования для обеих присадок была выбрана область концентраций от 0 до 1,4%. Эту область указали технологи на основании собственного опыта и неформализованных представлений до проведения планируемого эксперимента.

По предположению, поверхность в рассматриваемой области нелинейна, и простейшей здесь будет поверхность второго порядка. Уравнение регрессии имеет вид

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2.$$

Как вы, вероятно, помните из аналитической геометрии, поверхности второго порядка классифицируются на эллипсоиды, параболоиды, гиперболоиды, цилиндры в зависимости от знаков коэффициентов β_{11} и β_{22} и знака дискриминанта $4\beta_{11}\beta_{22} - \beta_{12}^2$ квадратичной формы, то есть вид поверхности определяется величинами и знаками трех последних коэффициентов.

Таблица 9

Условия и результаты экспериментов по окислению масла при использовании присадок типа *D* и *E*

№ эксперимента	Матрица планирования: x_1 — концентрация присадки D, x_2 — концентрация присадки E					Условия эксперимента в абсолютных единицах		Функция отклика — кислотное число
	x_1	x_2	$x_1 x_2$	x_1^2	x_2^2	концентрация D, %	концентрация E, %	y
1	+1	−1	−1	1	1	1,207	0,207	0,99
2	+1	+1	+1	1	1	1,207	1,207	0,76
3	−1	+1	−1	1	1	0,207	1,207	0,80
4	−1	−1	+1	1	1	0,207	0,207	0,76
5	−1,414	0	0	2	0	0	0,707	0,73
6	0	+1,414	0	0	2	0,707	1,414	1,14
7	+1,414	0	0	2	0	1,414	0,707	0,60
8	0	−1,414	0	0	2	0,707	0	1,10
9	0	0	0	0	0	0,707	0,707	0,83
10	0	0	0	0	0	0,707	0,707	0,78
11	0	0	0	0	0	0,707	0,707	0,72
12	0	0	0	0	0	0,707	0,707	0,85
13	0	0	0	0	0	0,707	0,707	0,74

Для нахождения оценок коэффициентов уравнения регрессии второго порядка использовался план на трех уровнях. В качестве уровней были взяты крайние значения в интервалах изменения переменных (они после преобразования имеют, как и раньше, значения ±1) и средняя точка, которая преобразуется в точку с координатой, равной нулю.

В таблице 9 приведены условия и результаты экспериментов, проведенных по плану второго порядка для двух переменных. После обработки результатов наблюдений и

отбрасывания незначимых коэффициентов регрессии получилась модель поверхности отклика в виде

$$y = 0{,}78 - 0{,}069x_1^2 + 0{,}158x_2^2.$$

Эта поверхность — гиперболический параболоид — имеет седлообразный вид, ее график вы видите на рис. 30. Движение вдоль направления, параллельного оси D, ведет к уменьшению кислотного числа y, и, следовательно, нам нужно двигаться в этом направлении. Дальнейшие шаги

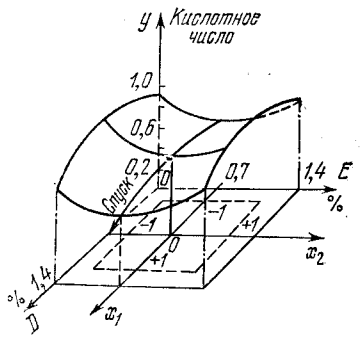


Рис. 30.

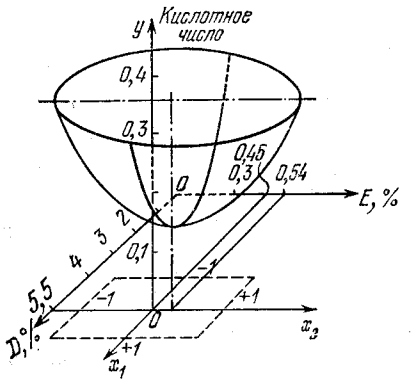


Рис. 31.

привели в точку q, соответствующую концентрациям 0,45% присадки E и 5,5% присадки D, в окрестности которой поверхность отклика, представленная на рис. 31, обладает четко выраженным минимумом.

Уравнение поверхности теперь имеет вид

$$y = 0{,}148 - 0{,}052x_2 + 0{,}093x_1^2 + 0{,}073x_2^2,$$

так что минимум достигается в точке, соответствующей концентрациям 0,54% присадки E и 5,5% присадки D. Критерий качества — кислотное число y здесь равно 0,14, что, как видите, значительно меньше, чем любой из результатов, получаемых в первой выбранной технологами области (последняя колонка в таблице 9), где минимальное значение кислотного числа было 0,6.

При использовании другой пары присадок, назовем их F и G, поверхность отклика имеет вид, представленный на рис. 32. Это также гиперболический параболоид,

но здесь поверхность доходит до плоскости $y=0$, и, следовательно, оптимальные точки располагаются вблизи линий уровня $y=0$ этой поверхности, например точка q_1. Не следует думать, будто на самом деле можно получить кислотное число y, равное нулю: нужно учитывать ошибку эксперимента и помнить, что все величины получены с точностью до этой ошибки.

Я не буду приводить другие примеры, ибо, надеюсь, вам уже понятно, что поверхности отклика в окрестности

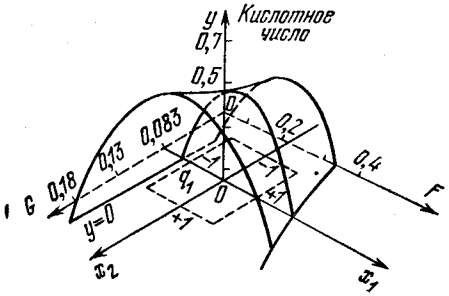


Рис. 32.

оптимальной точки могут быть весьма разнообразны, и их изучение, а затем и интерпретация полученных выводов в технологических терминах требуют и профессиональных знаний, и сноровки.

Итак, для нахождения оптимального соотношения концентраций присадок была применена шаговая стратегия движения по поверхности отклика вдоль направления, которое в каждой подобласти наиболее круто идет вниз. Впрочем, если бы нужно было найти не минимум, а максимум, то следовало бы двигаться по наиболее крутому пути вверх — здесь есть очевидная двойственность.

Рискну немного повторить сказанное выше. При шаговой стратегии сначала обследуется небольшая часть поверхности отклика, строится модель, адекватная этой части поверхности, проверяется гипотеза о достижении оптимума и принимается одно из решений «ДА», «НЕТ» или «МОЖЕТ БЫТЬ». При решении «ДА» поиск экстремума окончен, при решении «НЕТ» делается шаг по пути, наиболее круто идущему вниз (вверх), и все повторяется. При решении «МОЖЕТ БЫТЬ» нужно провести

дополнительные эксперименты для уточнения вида поверхности отклика. Это и есть перенесенная в эксперимент стратегия последовательного анализа Вальда — стратегия последовательного эксперимента — крупное достижение в теории эксперимента.

Все же в теории планирования эксперимента последовательная стратегия — далеко не единственное достижение. Применение многофакторного эксперимента, то есть решительный отказ от традиционного изменения факторов по одному при фиксации остальных факторов, оказывается не менее значительным достижением, чем стратегия последовательного эксперимента. В результате использования этих достижений существенно снижается количество необходимых экспериментов. Так, например, при проведении экспериментов для четырех факторов по пяти уровням надо провести $5^4=625$ опытов. Если же использовать один из видов оптимального планирования экспериментов (насыщенный D — оптимальный план), то те же результаты можно получить после 15 опытов.

В этом круге вопросов находится и организация такого порядка при проведении экспериментов, при котором не вырабатывались бы предрассудки и наступило бы избавление от систематических погрешностей, от которых при пассивном эксперименте избавиться трудно, а порой и невозможно. Здесь на помощь статистику, как это ни парадоксально, приходит случай, и через несколько страниц об этом будет подробный разговор.

Как достичь успеха

Среди 1093 патентов, выданных Бюро патентов США изобретателю Томасу Альва Эдисону, был патент № 223898 от 27 января 1880 года на лампу с угольной нитью накаливания. Три тысячи человек прибыли посмотреть на специальных поездах, заказанных Эдисоном, на сотни электрических лампочек, развешанных в его мастерской и над окрестными дорогами в Менло Парк (штат Нью-Джерси). Это было в последний день 1879 года. Но прежде чем была проведена эта победная демонстрация, Эдисон перепробовал шесть тысяч веществ, содержащих углерод, от обыкновенных швейных ниток, покрытых углем, до продуктов питания и смол. Лучшим оказался бамбук, из которого был сделан футляр японского пальмового веера.

Как вы понимаете, перепробовать шесть тысяч волокон — это же десятки тысяч опытов, и на эту титаническую работу ушло около двух лет. Думаю, если бы Эдисон владел теорией эксперимента, то количество опытов можно было бы значительно сократить, по-видимому, в несколько раз. Но во времена Эдисона еще не был развит факторный анализ, да и Эдисон не склонен был доверять такой не отвечающей ни его образованию, ни темпераменту науке, как математическая статистика.

Можно было бы рассмотреть с современных позиций эту работу Эдисона, но мне неизвестны ее детали,

Поэтому обращусь к более простому, так сказать, житейскому примеру.

Вся жизнь певицы зависит от ее успеха, ежедневного, проверяемого, лелеемого. Но вот наступает период конкурса, и ей нужен не просто успех, а победа. И она советуется с другом — математиком, какую выбрать арию и какой наряд. Не подумайте, будто от непоющего математика будет мало пользы. Но он, вместо восхищения тем черным элегантным платьем или чистотой исполнения арии Далилы, предлагает воспользоваться ежедневным успехом — коллективной оценкой зрителей. До конкурса есть еще некоторое время, и она успеет проверить себя на публике, причем проверить сознательно, целенаправленно. Для этого друг-математик предлагает применить метод, похожий на поиск наилучшей концентрации присадок к маслам, но, конечно, с заметными модификациями. Вот в чем его существо.

В отличие от проблемы поиска оптимальной концентрации присадок, факторы здесь не количественные, а качественные: подходящих нарядов у нее три, а хорошо подготовленных арий всего пять, и не укажешь какой-нибудь числовой переменной, изменяющейся непрерывно от черного шелкового платья до русского сарафана. Впрочем, непрерывность перехода от одного из платьев к другому не по существу дела: будь то фактор количественный или качественный, его возможные значения или варианты — они здесь называются уровнями — можно перенумеровать, что я ниже и сделал. Но когда фактор количественный, как концентрация, температура или вес, то порядок нумерации уровней не какой угодно — он отвечает возрастанию или убыванию реального значения уровня. Если же фактор качественный, то порядок нумерации уровней не играет никакой роли: все равно, как перенумеровать арии, наряды или материалы, которые испытывал Эдисон, создавая угольную лампочку. Поэтому качественные факторы требуют иного рассмотрения, чем количественные.

Теперь нужно подумать о результате наблюдений — надо как-то измерить успех, или, более точно, иметь возможность сопоставлять успех на разных выступлениях, знать, когда было лучше, когда хуже — нужно придумать критерий.

В качестве характеристики успеха можно выбрать продолжительность аплодисментов или их интенсивность, или количество вызовов к рампе. Но можно взять и какую-нибудь сравнительную качественную характеристику, скажем, назначить три категории: большой успех, средний и малый, или оценивать успех по пятибальной системе, как в школе. Конечно, при выборе такого субъективного критерия возникают новые проблемы: кого выбрать в качестве эксперта для выставления баллов, должен ли быть всегда один и тот же эксперт, или лучше опрашивать разных лиц, можно ли доверить актрисе оценку собственного успеха? Но не будем сейчас на этом останавливаться — пусть каким-то образом выбран критерий качества ее выступления.

Следует еще обратить внимание на зависимость успеха от аудитории: на концерте в консерватории и во Дворце культуры большого завода, в сельском клубе и в своем театре на концерте, заключающем торжественное заседание актива министерства будет разная реакция публики и на произведение, и на наряд, да и сама актриса волей-неволей будет вести себя по-разному. Поэтому мы рассмотрим три фактора: исполняемые арии, наряды и аудитории. Перечислю последовательно их уровни, закодировав их числами или буквами — это делается лишь для удобства различения факторов и ничего другого за обозначениями не подразумевается.

Арии

1. Ария Далилы («Самсон и Далила», Сен-Санс).
2. Ария Любаши («Царская невеста», Римский-Корсаков).
3. Частушки Варвары («Не только любовь», Р. Щедрин).
4. Хабанера («Кармен», Бизе).
5. Песнь Леля («Снегурочка», Римский-Корсаков).

Наряды

A. Черное узкое платье с глубоким декольте.
B. Синее бархатное платье с серебряной отделкой.
C. Русский сарафан с народным орнаментом.

Аудитории

α. Консерватория.
β. Дворец культуры завода.
γ. Сельский клуб.
δ. Концерт в театре для актива министерства.
ε. Дом культуры вуза.

Каждый из факторов что-то вносит в величину того параметра, которым мы измеряем успех, и каждый фактор (арии, наряды, аудитории) имеет разброс от уровня к уровню. Именно из-за этой неоднородности и приходится выбирать лучшую комбинацию из возможных.

Модель зависимости измеряемого параметра (y) от факторов можно записать аналогично регрессионной в виде суммы эффектов от каждого фактора и эффектов взаимодействий. Для иллюстрации запишу модель для двух переменных — арий и нарядов, пользуясь обычно применяемыми обозначениями:

$$y_{ij} = \mu + T_i + B_j + BT_{ij} + \varepsilon_{ij}.$$

Здесь μ — общий эффект во всех наблюдениях или истинное среднее совокупности, из которой взята наблюдаемая выборка, T_i соответствует эффекту от первого фактора на i-м уровне, то есть спетой арии, B_j — эффект от второго фактора на j-м уровне — от надетого наряда, BT_{ij} — эффект взаимодействия (может же актриса чувствовать себя неуютно, когда в русском сарафане поет Хабанеру!), y_{ij} — значение измеряемого параметра и, наконец, ε_{ij} — случайная ошибка в эксперименте. Однако эта модель имеет здесь другой, отличный от уравнения регрессии, смысл: она служит формулой, по которой вычисляется теоретическое значение измеряемого параметра y в отдельных (дискретных) точках — точках нашего плана.

А теперь отмечу самое главное: когда факторы действуют независимо, дисперсия измеряемого параметра равна сумме дисперсий слагаемых. Опираясь на это замечательное свойство дисперсии, можно провести дальнейший анализ, изучить вклады в дисперсию от каждого фактора, оценить относительную важность каждого из них, выбрать оптимальную их комбинацию. Применяемая здесь теория носит название дисперсионного анализа.

Я не буду, конечно, излагать дисперсионный анализ — как и всякий метод, он требует профессионализма, но попробую продемонстрировать некоторые идеи на примере с нашей актрисой.

Всех возможных комбинаций уровней, то бишь арий, нарядов и аудиторий будет $5 \times 3 \times 5 = 75$.

Для более наглядного представления нужно составить, например, для каждой из аудиторий таблицу, где по столбцу расположены арии — они закодированы цифрами, а по строке наряды — они закодированы прописными буквами (таблица 10).

Так как имеется пять аудиторий — уровней, то при проведении полного факторного эксперимента должно

Таблица 10

План полного факторного эксперимента для каждой из аудиторий

	A	*B*	*C*
1	*	*	*
2	*	*	*
3	*	*	*
4	*	*	*
5	*	*	*

быть пять таких таблиц. Вместо звездочек следует записать полученные результаты, то есть значения показателя y, полученные при исполнении арии и в наряде, соответствующих строке и столбцу, на пересечении которых стоит звездочка, и в таблице, которая относится к данной аудитории.

Но спеть 75 раз наша актриса не имеет возможности: до конкурса остался месяц, и она успеет в лучшем случае спеть в пять раз меньше. Вот тут-то друг-математик и подсказывает стратегию. Она состоит в том, чтобы каждая ария была исполнена хоть один раз в каждом из нарядов, и каждый наряд был показан хотя бы раз в каждой из аудиторий. Не сбалансированы только арии и аудитории, и поэтому план называется частично сбалансированным (таблица 11).

Таблица 11

Частично сбалансированный неполноблочный план

Наряды / Арии	*A*	*B*	*C*
1	α	β	γ
2	β	γ	δ
3	γ	δ	ε
4	δ	ε	α
5	ε	α	β

Каждая клетка схемы предписывает произвести эксперимент, в котором реализуется сочетание из наряда (номера столбца), арии (номера строки) и аудитории (греческой буквы, стоящей на пересечении столбца и строки). Вот в этом плане и содержится 15 экспериментов.

После проведения эксперимента в клетку заносится результат, то есть полученное значение выбранного критерия, и затем производится обработка всей таблицы.

Не буду рассказывать о всей математической «кухне» — это надо уже изучать по учебнику. Но хочу подчеркнуть: дисперсионный анализ — весьма эффективный метод для сравнения подобных сочетаний или композиций и выбора наилучшей. Скажем, приведенный план был нами использован все в той же задаче о выборе присадок к маслам, но уже не при определении оптимальных концентраций, а при отборе наилучшей композиции присадок в ситуации, где было пять антиокислительных присадок, три противоизносные и пять антикоррозионных. Без методов математической статистики пришлось бы провести испытания всех 75 возможных комбинаций.

Однако применение методов планирования экспериментов, именно плана, приведенного в таблице 11, при отборе лучшей композиции дало возможность не только сократить число опытов до 15, причем практически без ущерба для полноты дальнейших выводов, но и позволило выявить ряд важных технологических фактов о поведении присадок, эффектах их взаимодействия, в том числе и такие факты, которые без математической обработки установить было бы невозможно.

Дисперсионный анализ широко применяется в психологии, биологии, химии — всюду, где встречаются задачи с качественными факторами. Впрочем, дисперсионный анализ применяется и при количественных факторах, но о них у нас уже была речь, хотя я рассказывал выше о других методах.

Случай — помощник

Покупая лотерейный билет, перебегая дорогу перед носом автомобиля или выходя замуж после двухмесячного знакомства, люди полагаются на счастливый случай.

Есть и другая ситуация, где без вмешательства случая либо вообще нет явления, либо именно случай и дает возможность уверенно решать поставленные задачи.

С подобной ситуацией, когда случай служит единственным механизмом явления, я столкнулся при изучении ряда вопросов функционирования радиорелейных линий в ультракоротковолновом диапазоне, то есть в диапазоне электромагнитных излучений с длиной волн не больше 10 метров. Если длинные и средние волны, на которых ведут передачи радиовещательные станции, огибают поверхность Земли, а короткие отражаются от ионосферы, то волны ультракоротковолнового диапазона проникают сквозь ионосферу и почти не огибают выпуклости на поверхности Земли. Распространение ультракоротковолновых излучений практически происходит в пределах прямой видимости, подобно лучам света. В то же время ультракоротковолновый диапазон весьма привлекателен по целому ряду физических свойств и технических характеристик.

Однако, несмотря на установленные факты прямолинейного распространения излучений ультракоротковолнового диапазона, были экспериментально обнаружены и

какие-то аномалии: например, в СССР неожиданно были приняты передачи бельгийского телевидения.

Не буду рассказывать историю вопроса и радиофизические детали, а поясню грубо механизм, объясняющий, как это могло произойти.

Нижние слои атмосферы, называемые тропосферой, постоянно находятся в турбулентном режиме, то есть наблюдается беспорядочное вихревое движение воздуха. Вихри можно и увидеть, если понаблюдать за дымом фабричных труб: дым обычно клубится по сложным извилистым весьма причудливым траекториям. Турбулентность возникает по многим причинам: ветры, воздушные течения, неравномерность нагревания Солнцем различных участков Земли и т. д.

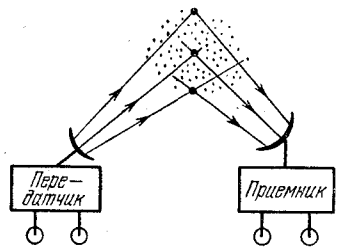


Рис. 33.

Турбулентное движение приводит к случайным изменениям плотности, температуры, влажности воздуха, что в свою очередь приводит к флуктуациям показателя преломления воздуха и его диэлектрической проницаемости.

В качестве модели такой турбулентной тропосферы рассматривают совокупность центров рассеяния. Центр рассеяния можно представлять себе как шарик, и множество таких шариков, расположенных в пространстве случайно, подобно изюминкам в калорийной булочке, представляет собой обсуждаемую модель тропосферы. Когда на область турбулентной тропосферы поступает электромагнитная энергия от передатчика, то образуется поток рассеянной во все стороны энергии. Конечно, основная часть энергии направлена вдоль первоначального направления волны, но все же некоторая не очень значительная, но достаточная для воспроизведения часть энергии, отражаясь от центров рассеяния, поступает на вход приемного устройства, антенная часть которого может находиться в области тени передатчика. На рис. 33 и представлена эта ситуация.

Сейчас созданы радиорелейные линии передачи, использующие дальнее тропосферное распространение ультракоротких волн, и случайный механизм рассеяния волн в турбулентной тропосфере является единственным

механизмом, обеспечивающим функционирование линии. Если бы вдруг тропосфера «затихла», прекратились бы турбулентные флуктуации, то немедленно прекратила бы прием и радиорелейная линия — сигналы перестали бы поступать на приемную антенну. Так случайный механизм оказывается основой метода передачи сигналов.

Построению математической модели, связывающей функцию отклика с факторами, мешают неизвестные, а то и известные переменные — состояния объекта, которые трудно поддаются учету или контролю.

Трудность — еще не невозможность... Но как же избавиться от этих переменных, мешающих изучению нужных зависимостей? Кажется ясным: надо отделить изучаемую зависимость от мешающих переменных, и если такое разделение нельзя произвести абсолютно точно, то следует это сделать с допустимой точностью...

Однако для большинства сложных систем разделить «полезные» и «мешающие» переменные нельзя. Скажем, вы отравились — съели что-то несвежее. Казалось бы, должен на это реагировать желудочно-кишечный тракт. Но, кроме болей в области живота, у вас наблюдается повышение пульса и температуры — реакция сердечно-сосудистой системы, изменилось настроение — реакция нервной системы. Все эти реакции в организме тесно связаны, и разделить их нельзя. Впрочем, и в значительно более простых, чем живая природа, технологических процессах не удается разделить подобные переменные или их функции.

Но даже и тогда, когда состояния объекта и входы независимы, когда их можно разделить и изучать поочередно, учесть все варианты практически невозможно. Давайте оценим размеры бедствия. Скажем, в процессе первичной переработки нефти производится разделение нефти на несколько фракций: бензины разных марок, реактивное и дизельное топливо, газойль, масла нескольких марок, гудрон, словом, подчас более десятка фракций. Даже если бы каждая из фракций определялась всего двумя числами — верхним и нижним уровнем, то вариантов было бы порядка 2^{10}, то есть более тысячи. На самом же деле каждый из компонентов описывается многими числами, и поэтому количество вариантов колоссально возрастает. Но, главное, состояния объекта — это более сотни регулируемых или управляемых параметров: температуры и давления в разных точках огромной установки, расходы сырья и пара

и многое другое. Даже если бы эти величины могли принимать значения лишь на двух уровнях, то вариантов было бы порядка 2^{100}, что более 10^{30}, то есть число с тридцатью значащими цифрами. С подобными огромными числами никакие математические машины, даже самые быстродействующие, справиться не могут.

Таким образом, по существу, попытка избавиться от мешающих, ненужных переменных за счет их скрупулезного учета — это происки все того же дьявола, который толкает исследователя на бесперспективный путь однофакторного эксперимента, нашептывая ему: «зафиксируй значения всех остальных переменных». Просто здесь дьявол принарядился, поменяв необработанную овечью шкуру на более модную дубленку.

Конечно, такое положение не облегчает жизнь экспериментатору, а вопрос об избавлении от мешающих переменных остается открытым.

Все же выход есть, и он совсем не банален: вместо скрупулезного учета изменений, мешающих в исследовании переменных, надо широко воспользоваться случаем, поставить его себе на службу.

Вот в чем здесь суть дела. Для примера займем место рядом с исследователем — психологом. Он должен выяснить время, затрачиваемое детьми в четвертом классе на решение арифметических задач определенного типа. Для испытаний выделена группа из пяти девочек и пяти мальчиков.

На основании предыдущего опыта нельзя сделать заключение о меньших или больших успехах мальчиков по сравнению с девочками при решении таких задач. В каком порядке нужно подвергать испытанию детей в этой группе?

Можно, например, из галантности сначала пропустить девочек. Но перед испытаниями в классе был урок физкультуры, причем девочки играли в баскетбол, а мальчикам объясняли правила самоконтроля, так что девочки и мальчики в неравном положении: у девочек может сказаться усталость. Кроме того, выбранные десять ребят не одинаковы. Во-первых, они разного возраста: два мальчика и четыре девочки старше десяти лет, а остальные — младше. Один из мальчиков — второгодник, а две девочки — отличницы. В каком же порядке подвергать испытанию эту десятку?

Ответ здесь один: нужно обеспечить случайный порядок проверки. Именно за счет случайного

порядка в значительной степени усреднятся эффекты, возникающие из-за усталости части детей, разброса их возрастов или вследствие различия в подготовке и добросовестности.

Аналогичные проблемы постоянно возникают в экспериментальной биологии, медицине, когда на группе белых мышей или морских свинок производят проверку действия каких-либо ядов, облучений или лекарств. Биологи обычно делят животных на две группы — испытуемых и контрольных. Скажем, имеется сорок белых мышей. Как следует разделить их на две группы, сколько особей взять в каждую из групп, в каком порядке их подвергать все возрастающим дозам облучения? Если вы посмотрите на этих безобидных зверьков, то вам они покажутся более или менее одинаковыми, и представляется вполне допустимым разделить их, как придется, например, поделить группу пополам, а затем из одной подгруппы брать подряд, не задумываясь...

Именно на этом бездумном пути вырабатываются предрассудки. Неоднократно были установлены факты: в ситуации, когда человеку кажется, будто он поступает как попало, по воле случая, его действия оказываются целенаправленными. Здесь играют роль какие-то еще мало изученные механизмы подсознательной деятельности. Биолог подчас отбирает для опыта подгруппу более слабых животных, когда ему хочется доказать эффективность яда, или, наоборот, более сильных, когда он доказывает эффективность противоядия. Заметьте, я не говорю о сознательном обмане — вовсе нет! Здесь работают подсознательные механизмы.

Мой близкий друг лет двадцать пять назад обсуждал со мной потрясший его факт, и мы вместе пытались тогда найти ему объяснение. Он поручил девушкам-вычислителям произвести расчеты таблиц стрельб для одного вида оружия, которым он тогда занимался. Ему представили таблицу, где было десять тысяч пятизначных чисел. Он полистал странички, выбрал наугад одно число, сам провел вычисления и обнаружил ошибку. Тогда он взял другое число, тоже наугад, и после вычислений вновь обнаружил ошибку. Страшно рассердившись, он накричал на вычислителей и заставил пересчитать всю таблицу вновь, теперь уже «в две руки», то есть проверяя результаты вычислений по этапам, когда их производят сразу два вычислителя. Все остальные 9998 чисел оказались верными!

Он не сомневался, что выбрал два числа «наугад», то есть совсем случайно, и ему представлялась равновероятной возможность выбора любого из этих десяти тысяч чисел. При этом вероятность выбора именно этих двух дефектных чисел есть 2/10 000, то есть весьма маленькое число, и по принципу практической уверенности следовало бы ожидать, что подобное событие не произойдет.

Вот такое обстоятельство и вызывало серьезное удивление и недоверие. Однако сегодня ясно: мой друг был человек опытный и выделил числа, которые хотя и не очень заметно, но все же уклонялись от значений, которые должны были бы быть. Но сделал это подсознательно, не отдавая себе отчета в мотивах своих поступков.

Таким образом, его выбор был не случаен, или не совсем случаен, и уж, конечно, нельзя было считать извлечение им чисел для контроля из таблицы равновероятными и независимыми событиями.

Теперь настало время объяснить, как же обеспечить действительно случайный отбор детей, чисел или морских свинок, выбор, обеспечивающий независимость от воли экспериментатора порядка обследования объектов или проведения опытов. Для этого используют таблицы случайных чисел. Правило составления таких таблиц весьма просто. Надо взять десять одинаковых шариков, написать на них числа от нуля до девяти и положить их в мешок. Тщательно перемешав, вынуть один, записать полученное число, а шар возратить в мешок. Вновь тщательно перемешать шары и повторить процедуру — так получится второе число. Повторяя много раз эту процедуру, мы и составим таблицу случайных чисел. Их можно записать в колонки пятизначных чисел или шестизначных, или как вам будет угодно. Нечто подобное вы можете наблюдать по телевизору при розыгрыше спортлото, только здесь шары не возвращаются — в спортлото нельзя допустить повторение чисел.

Конечно, на самом деле никто так не делает, но все же многие реальные способы, использующие электронные быстродействующие математические машины, по существу, моделируют описанную процедуру.

Когда нужно установить случайный порядок при проведении эксперимента, следует сначала перенумеровать объекты (например, морских свинок), а затем разбить на группы с помощью таблицы случайных чисел. Скажем, у вас сорок морских свинок. Вы приписываете им каким-то образом номера. А затем отбираете в контрольную группу первые двадцать, двузначные номера которых встретятся среди первых пар таблицы случайных чисел (номера, большие сорока, отбрасываются, ибо для них нет соответствующих особей). Также отбирается и испытуемая группа, причем порядок в ней диктуют последующие числа все той же таблицы.

В задаче выбора оптимального соотношения присадок к маслам последовательность экспериментов по той же логике должна быть случайной. И певица в борьбе за успех должна выбирать порядок своих выступлений (арий, нарядов и аудиторий) также по воле случая: выбор клеток в матрице частично сбалансированного неполноблочного плана (таблица 11 в предыдущем разделе) следует производить по таблице случайных чисел.

Таким образом, случайный порядок отбора значений факторов служит надежным средством избавления от предрассудков, которые могут выработаться при экспериментальном изучении процессов или явлений. И статистик сегодня, составляя порядок проведения экспериментов, обязательно пользуется случайным порядком, или, как говорят специалисты, рандомизацией (от английского random — случай), используя случай как надежного помощника.

Реплики под занавес

Давайте пролистаем книжку. О чем же она написана? Здесь обсуждались статистическая проверка гипотез и последовательный анализ, теория риска и построение математических моделей, идентификация и прогнозирование, теория пассивного и активного эксперимента. Этот перечень научных терминов создает впечатление, будто книжка посвящена довольно специальным вопросам. Но мне представляется ее главное содержание иным. Будь вы, читатель, школьник или станочник, биолог или радист, экономист или археолог, будь вы даже администратор на любом уровне нашей сложной иерархической системы управления, в своей работе и повседневной жизни вы оказываетесь то экспериментатором, то наблюдателем, то лицом, принимающим решения. Ставить свои задачи, проводить наблюдения или эксперименты и принимать решения вам приходится в сложной обстановке, в тумане случайности, подчас довольно густом. И поэтому небесполезны идеи и методы статистической теории управления и эксперимента.

Конечно, методы планирования эксперимента обеспечивают экономию, и подчас значительную, количества необходимых опытов. Но главное не в этом.

Когда обсуждается поездка из Москвы во Владивосток или даже в Крым и сравнивается поезд с самолетом, то почему-то говорят об экономии времени. Мне представляется это недоразумением или, если угодно, традиционным непониманием преимуществ самолета по сравнению с поездом. Важнее совсем другое: если вы летите, то вам не грозит длительное пребывание в обществе неизвестных попутчиков, которые могут оказаться и милыми людьми, и они не будут курить в купе или сопротивляться открыванию окна, когда вам уже совершенно необходим кислород, не будут требовать, чтобы вы бросили книгу, слезли с полки и выпили стакан водки за здоровье их детей, не будут охранять от вас годовалых близнецов, страдающих расстройством желудка, никто из попутчиков не будет храпеть всю ночь или сушить носки как раз там, где у вас лежит вареная курица, и вам не нужно жевать эту чертову птицу, липкую и жесткую, а затем стоять в очереди у туалета, дабы помыть не только руки, но и уши, и, возможно, у вас не будет обострения холецистита, колита или диабета после этого сухохолодоедения. Впрочем, букет болезней может бурно расцвести в результате посещения вагона-ресторана, где вы получите еще и порцию эмоций при обсуждении с официанткой вопроса о том, кто же кого здесь обслуживает. Словом, купив билет на самолет, вы с большей вероятностью избавляете себя от участия в грустных историях из жизни. Этому посвящены сотни фельетонов, мораль которых состоит в реальной возможности избежать всех подобных неприятностей, если бы отдельно взятые стрелочники и их коллеги на железной дороге вели себя в соответствии, а пассажиры соблюдали бы...

Так и в теории эксперимента — главное достижение вовсе не в сокращении числа необходимых опытов. Сознательное планирование эксперимента влечет за собой необходимость четкого логического анализа всей ситуации от продумывания исходных посылок, анализа априорной информации и выбора адекватной модели до использования рандомизации, последовательной стратегии эксперимента, оптимального выбора точек в факторном пространстве, разумной, обоснованной интерпретации результатов статистической обработки, причем представленной в компактной форме, удобной и для публикации, и для дальнейшего использования.

Искусство исследователя в равной мере проявляется в выборе модели — хорошей, простой, доступной и адекватной, и эффективного метода ее использования. А коль скоро нужно учитывать случайный характер обстановки,

то и в выборе эффективного статистического метода. Поэтому мало, хотя и необходимо, иметь статистическое образование и уметь им пользоваться, нужно еще «изнутри» понимать свою задачу и окружающие ее проблемы. И то, и другое сразу не всегда легко дается, но можно прибегнуть к кооперации, совместной работе экспериментатора со статистиком.

Надеюсь, мне удалось показать идейную и методическую связь статистической теории управления и теории эксперимента, которые я и объединил одним названием, и их значимость для вас, ваших коллег и знакомых, да и незнакомых тоже.

Словом, я надеюсь, вы хотя бы немного прониклись идеями статистической теории управления и эксперимента и уже не скажете «НЕТ» этой теории. И если вам кажется еще рановато сказать «ДА» и с завтрашнего дня систематически применять в своей деятельности методы, о которых шла речь в книжке, то, небось, пора выбрать стратегию «МОЖЕТ БЫТЬ...» и приступить к их серьезному изучению.

Желаю вам успехов на этом пути.

Оглавление