



АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ

Серия «Наука и технический прогресс»

Ю. И. Гильдерман

Закон и случай

Ответственный редактор
доктор биологических наук
В. А. РАТНЕР



НОВОСИБИРСК
«НАУКА»
СИБИРСКОЕ ОТДЕЛЕНИЕ
1991

ББК 22.171
Г47
УДК 16.2.9—15.1

Рецензенты

доктор биологических наук *А. Б. Рубин*
кандидат физико-математических наук *Е. М. Сорокина*

Гильдерман Ю. И.

Г47 Закон и случай.— Новосибирск: Наука. Сиб. отд-ние, 1991.— 200 с. (Серия «Наука и технический прогресс»).

ISBN 5—02—029907—3.

Почему нельзя проводить безлимитные шахматные матчи на первенство мира? Как улучшить правила игры «Что? Где? Когда?» Сколько нужно иметь детей, чтобы хотя бы один из них был мальчик? Есть ли шансы сбить самолет из ружья? На эти и другие вопросы вы найдете ответ в нашей книге. В ней рассказывается о классических методах теории вероятностей и их приложениях.

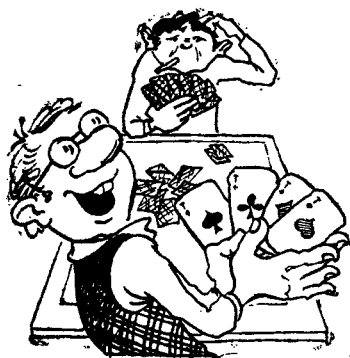
Книга рассчитана на специалистов по теории вероятностей и широкий круг читателей.

Г $\frac{1602110000-170}{054(02)-91}$ 58—91 НП

ББК 22.171

ISBN 5—02—029907—3

© Издательство «Наука», 1991



Слова «случай», «случайность», «случайпо» едва ли не самые употребительные в любом языке. Часто они окружены романтическим ореолом. Говорят, например, «Его Величество Случай». С этими словами связывают мистические понятия Рока, Судьбы. Их окружают туманом непредсказуемости, неопределенности. Их противопоставляют ясной и четкой информации, строгому логическому развитию событий, защищенному от всякой случайности.

Однако так ли уж велика пропасть между случайным и неслучайным? Конечно, каждая отдельно взятая молекула под случайными ударами соседних молекул совершает хаотическое, непредсказуемое в деталях движение, которое называют броуновским. Но вся масса молекул подчиняется вполне строгому закону диффузии, который можно описать дифференциальным уравнением. Пользуясь этим законом, нетрудно установить скорость диффузии, заранее вычислить плотность исследуемого вещества в указанной точке сосуда, найти другие вполне определенные, т. е. неслучайные, характеристики процесса.

Точно так же никто не возьмется предсказать судьбу отдельного человека. Никто не скажет, женится ли он, сколько у него будет детей, будет ли он болеть туберкулезом или страдать от повышенного кровяного давления, до какого возраста он доживет. Однако в любой стране или даже большом городе известно, сколько в среднем мужчин из каждой тысячи двадцатипятилетних вступает в брак, какова рождаемость и смертность, сколько в среднем детей имеет каждая семья, каковы показатели заболеваемости туберкуле-

зом или гипертонией и т. д. и т. п. Все это — более или менее твердо установленные на данный период характеристики страны или города. Во многом потерявшие черты случайности, они публикуются в ежегодных справочниках. Опираясь на них, планируют экономическое и социальное развитие, разрабатывают дорогостоящие программы преобразования регионов. Знание этих уже неслучайных характеристик и служит защитой от непредвиденного, непредсказуемого, туманного и неопределенного. Таким образом, случайность, когда она проявляется в поведении не одного объекта, а многих сотен или даже тысяч похожих объектов, обнаруживает черты закономерности. Путь, которым необходимость идет к цели, — говорят философы, — вымощен бесчисленным множеством случайностей.

Вот эту необходимость, т. е. закон, строгий рисунок которого соткан из многих и многих случайностей, и старается установить наука, названная в конце концов теорией вероятностей.

На парадоксальность сочетания слов «теория» и «вероятность» обратил внимание один из самых первых исследователей случайного — французский философ, писатель, физик и математик Блез Паскаль (1623—1662). Размышляя над двумя задачами о стратегии игры в кости, предложенными ему придворным де Мере, обсуждая решения этих задач в переписке со знаменитым математиком Пьером Ферма (1601—1665), Паскаль пришел к выводу о существовании строгих законов, описывающих случайные явления. В письме в Парижскую академию наук в 1654 году он писал: «И так как строгость научных доказательств сочетается здесь с неопределенностью случайности, то трактат, где эти два предмета, с виду противоположные, как бы примирились друг с другом, может по праву претендовать на ошеломляющее название „Математика случая“».

К сожалению, болезнь и ранняя смерть помешали Паскалю написать задуманный трактат, и о полученных им результатах можно судить лишь по его письмам и свидетельствам современников. Материалом для исследований Паскаля по теории вероятностей послужила азартная игра в кости. Но почти за сто лет до рождения Паскаля «Книгу об игре в кости» написал итальянский математик Дж. Кардано (1501—1576).

Сочинение «О выходе очков при игре в кости» имеется и в трудах Г. Галилея (1564—1642). Такое внимание выдающихся математиков к азартным играм легко объяснить. Как мы уже отмечали, законы теории вероятностей можно обнаружить только при большом числе однотипных случайных явлений, при многократном повторении опыта. Азартные игры в карты и кости, чрезвычайно широко распространенные в Европе в XVI—XVII веках, как раз и явились теми простыми и многократно повторяющимися опытами, опираясь на которые можно было нащупать и проверить предполагаемую закономерность. Поэтому и первый более или менее полный трактат по теории вероятностей, написанный в 1657 году голландским математиком, механиком и астрономом Христианом Гюйгенсом (1629—1695), все еще назывался «О расчетах при игре в кости или о расчетах при азартной игре». Но уже тогда Гюйгенс подчеркнул, что дело не в азартных играх, а в том, что «закладывается основа очень интересной и глубокой теории».

История очень быстро подтвердила это предсказание Гюйгенса. Развитие производства, возникшие задачи страхования и демографии, усложнение физических экспериментов и астрономических наблюдений, связанных с измерениями, резко расширили границы приложений теории вероятностей, привлекли к ней внимание новых поколений математиков. Целая плеяда звезд — Я. Бернулли (1654—1705), А. Муавр (1667—1754), Д. Бернулли (1700—1782), Т. Байес (1702—1761), Л. Эйлер (1707—1783), Ж. Бюффон (1707—1788), Ж. Д'Аламбер (1717—1783), П. Лаплас (1749—1827), С. Пуассон (1781—1840), К. Гаусс (1777—1855) — блок за блоком уложила фундамент, а затем возвела и само здание «математики случая».

Современное развитие теории вероятностей связано с именами П. Л. Чебышева (1821—1894), А. А. Маркова (1856—1922), А. М. Ляпунова (1857—1918), К. Пирсона (1857—1936), П. Леви (1886—1971), Р. А. Фишера (1890—1962), А. Я. Хинчина (1894—1959), В. И. Гливенко (1897—1940), А. Н. Колмогорова (1903—1987).

«Замечательно, что наука, которая начала с рассмотрения азартных игр, обещает стать наиболее важным объектом человеческого знания... Ведь большей

частью важнейшие жизненные вопросы являются на самом деле лишь задачами из теории вероятностей». Эти слова также стали пророческими. Но удивительно, что они принадлежат П. Лапласу — автору не только «Аналитической теории вероятностей», но и так называемого принципа классического детерминизма (от латинского слова *determino*, что означает «определяю»). По этому принципу весь мир от молекулы до планет и живой природы устроен как гигантский часовой механизм. Только в силу того, — утверждали приверженцы этого принципа, — что пока не до конца известны все связи между причинами и следствиями, приходится пользоваться иногда теорией вероятностей. Но как только эти причины будут установлены, движение и молекул, и планет, и жизнь человека можно будет описать детерминистически, т. е. определенно, без уступок случаю, например, с помощью большой системы дифференциальных уравнений.

Конечно, заманчиво было бы построить такую математическую модель развития человека и человеческого общества, которая позволила бы с точностью до секунды предсказывать наступление тех или иных событий в обществе, как астрономы предсказывают затмение солнца или появление комет. Однако мир, и живой, и неживой, оказался намного сложнее, чем это представлялось 150 лет назад. Некоторые явления позволяют выделить основные черты их механизма, описать эти черты с помощью таких средств, как дифференциальные уравнения (связывающие скорости изменения изучаемых величин с самими этими величинами), построить, как говорят, их детерминистическую модель. Такие модели широко распространены в механике, астрономии, физике, технике. Имеются детерминистические модели в химии и биологии. С их помощью описывается движение крупных тел, течение химических реакций, развитие в целом больших популяций и т. п. Имеются и такие объекты, как, например, небольшие популяции, которые в равной степени проявляют и детерминистические, и вероятностные черты.

XX век с его бурным развитием массового автоматизированного производства, прогрессом науки и эксперимента, лавинообразным ростом народонаселения и возникшими в связи с этим проблемами демографии, медицины, транспорта и связи поставил исследователей перед необходимостью изучения таких явлений, кото-

рые просто невозможно представить в духе детерминизма как жесткую цепь следующих друг за другом причин и следствий. Точнее говоря, микропричины и следствий, составляющих ткань изучаемого явления, слишком много и отдать чему-то предпочтение, заранее выделить что-то бесспорно главное, а что-то пренебрежимо второстепенное невозможно. Отдельные проявления микропричин носят хаотический непредсказуемый характер. Нельзя предугадать, что в эту минуту откажет эта деталь, что в этот день заболит этот человек, что в эту секунду произойдет распад этого атома. Приходится изучать всю массу отдельных проявлений. Здесь-то и используется теория вероятностей, которая «позволяет найти порядок и закономерность там, где классический детерминистический подход оказывается бессильным». Эта характеристика принадлежит известному советскому математику и педагогу Б. В. Гнеденко, одну из популярных книг которого мы приводим в рекомендуемом списке литературы.

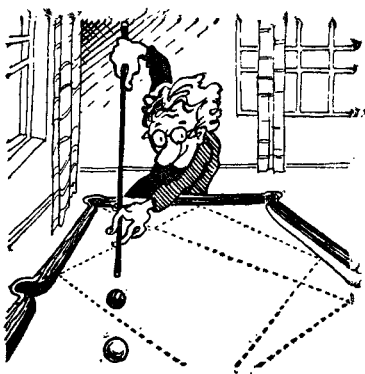
Сейчас невозможно указать ни одной области человеческой деятельности — от заводского производства до литературоведения, — где бы не применялись вероятностные исследования. Поэтому теорию вероятностей называют «знаменем математики XX века». Говорят также о «стохастической революции в сознании». В современном языке «стохастический» означает случайный, вероятностный. Но полезно помнить и о первоначальном смысле этого слова. *Stochastikós* по-гречески означает «умеющий угадывать», т. е. тот, кому ведомы законы случайного.

Что же касается химии и биологии, то изучение объектов этих наук — реакций с участием миллионов однотипных молекул, популяций, состоящих из тысяч особей, поколений, сменяющих друг друга на протяжении миллионов лет, и т. п. — привлекло математиков-вероятностников, стимулировало развитие теории вероятностей и математической статистики. Не случайно основы современных статистических методов были заложены К. Пирсоном и Р. А. Фишером, известными в науке не только как математики, но и как биологи.

В предлагаемой книге рассказывается о некоторых, давно уже ставших классическими, методах и приемах теории вероятностей и о их применении в естествознании и других областях знания человека.

Случайные события

1. Испытания, события, вероятности



Классическим примером, который служит иллюстрацией ко многим задачам теории вероятностей, является подбрасывание монеты. Мы не можем определенно сказать, какой стороной — гербом или номиналом — упадет монета. На исход подбрасывания влияет очень много разнообразных и трудно учитываемых факторов — сила броска, скорость вращения, сопротивление воздуха, небольшая неоднородность тела монеты, характер рисунка на ней, упругие свойства монеты и поверхности, на которую она падает и т. д. и т. п. Описать все это строго и однозначно с помощью, например, дифференциальных уравнений или другими детерминистскими методами, надеясь точно указать, при каких условиях монета упадет гербом, а при каких — номиналом, совершенно невозможно. Но если задачу невозможно решить, то, наверное, и пытаться не стоит. Лучше изменить ее постановку так, чтобы задача не потеряла практического смысла, но допускала разумное решение. Зададимся, например, вопросом, как часто монета будет падать гербом. Такой вопрос имеет ясный практический смысл, и каждый на него готов ответить. Если монета более или менее однородна и бросается много раз, то примерно в половине случаев она будет падать гербом. Этот вывод формулируют кратко: вероятность выпадения герба при данном бросании равна $1/2$.

Еще один пример. Имеется ящик с небольшим отверстием сверху. Будем называть его урной. Положим в урну 10 шаров. Шары одинаковые на ощупь, по части из них белые, а остальные — черные. Тщательно перемешаем шары и будем вынимать по одному. Есть ли

у нас полная уверенность в том, что, например, при первом вынимании будет извлечен белый шар? Разумеется, нет. Лишь в двух крайних случаях мы можем точно предсказать цвет вынимаемого шара. Если все шары черные, то появление белого шара невозможно, и, наоборот, если все шары белые, то каждый вынутый шар будет белым. В остальных случаях у нас есть лишь некоторая *доля уверенности* в том, что шар будет белым, и эта доля тем больше, чем большую долю составляют белые шары в урне.

Совершенно естественно возникает задача описать эту долю уверенности, пайти объективную числовую характеристику для нее. Такой характеристикой и является вероятность.

В связи с этим надо сказать, что слово «вероятность» в бытовой речи часто употребляется в ином смысле. Так, например, знатоки футбола, опираясь на информацию об играх данной команды в прошлые сезоны, на разнообразные сведения о том, что кто сказал, кто в какой форме, какое поле, какая погода и т. д. и т. п., предсказывают с некоторой вероятностью исход матча и даже счет. Иногда сведений так много, что для их обработки привлекают ЭВМ. Не подвергая сомнению интерес подобных исследований, тем не менее, отметим, что они не входят в круг забот теории вероятностей. Футбольный матч классных команд — это уникальное, неповторяющееся событие. С данными составами и тренерами в данном чемпионате оно происходит один-два раза. Нельзя устроить достаточно длинную серию подобных матчей с неизменными составами при неизменных условиях. Известно, что даже к концу матча команды совсем не похожи на те, какими они были вначале. В отличие от этого, как мы уже говорили, теория вероятностей занимается сериями ординарных массовых испытаний, которые можно повторять в примерно одинаковых условиях много раз. Многократные подбрасывания монеты или игральной кости, стрельба по мишени, изготовление массовых деталей на заводе, рождение детей в стране или крупном городе, вызовы абонентов на телефонной станции, популяции, состоящие из тысяч однотипных особей, многократно повторяемые скрещивания генотипов, громадные молекулы, состоящие из сотен и тысяч мономеров, серии однотипных химических реакций — вот примеры

объектов и явлений, которые изучают с помощью теории вероятностей.

Приступая к построению математической теории, мы, как обычно, должны ввести исходные, первичные понятия. В теории вероятностей такими понятиями являются *испытание* и *исход испытания*. Сошедшую с конвейера деталь испытывают с целью выяснить наличие отклонений от стандарта. Обнаружение отклонения или заключение о его отсутствии — исход испытания. При скрещивании генотипов Aa и Aa могут появиться генотипы AA , Aa и aa . Таким образом, скрещивание — это испытание, а появление в потомстве одного из указанных генотипов — это три возможных исхода испытания. Заметим, что испытанием в данном случае может быть не скрещивание (которое часто, например в природе, происходит без участия испытателя), а лишь наблюдение, анализ наугад взятой особи из потомства, подобно тому как испытывают наугад взятую деталь.

Точно так же бросание монеты — испытание, выпадение герба или номинала — один из двух возможных исходов испытания. Вместо слов «исход испытания» будем говорить *случайное событие* или просто событие.

С каждым событием A мы хотим связать некоторое число $P(A)$, заключенное между нулем и единицей, которое будем называть *вероятностью этого события*. Это число должно характеризовать частоту наступления события A при многократном повторении испытания.

Например, вероятность выпадения герба равна $1/2$. Это значит, как мы уже отмечали, что при достаточно большом числе испытаний (брошенных монеты) примерно в половине случаев выпадет герб.

Аналогично, если испытанием является бросание игральной кости (однородного кубика, на каждой грани которого отмечено определенное для этой грани число от 1 до 6), то естественно считать, что выпадение загаданной грани равно $1/6$. Это означает, что при достаточно большом числе испытаний примерно в одной шестой случаев выпадет указанная грань.

Такой же смысл имеет вероятность $P(A)$ в общем случае. Она характеризует частоту, повторяемость события A в достаточно большей серии испытаний. Если вероятность наличия дефекта в детали равна 0,013, то мы не можем сказать, будет ли наугад взятая деталь

исправной или дефектной. Но мы можем утверждать, что в каждой партии из 1000 штук примерно 13 деталей будут бракованными. Точно так же, если каким-то образом установлено, что вероятность появления в потомстве альбиноса равна 0,25, то это значит, что примерно четверть потомства будет состоять из альбиносов.

Если наступление события A столь же возможно, как и наступление события B , то такие события будем называть *равновозможными* и приписывать им равные вероятности. Это значит, что при повторении испытаний событие A наступает так же часто, как и событие B . Например, выпадение герба и номинала при подбрасывании монеты — равновозможные события. Аналогично выпадение различных граней при бросании игральной кости — равновозможные события.

Невозможному событию будем приписывать вероятность, равную нулю. Например, появление белого шара, если в урне только черные, — невозможное событие. Его вероятность равна нулю.

В противоположность невозможному событие может быть *достоверным*. Например, появление белого шара, если в урне только белые, — достоверное событие. Такому событию естественно приписать вероятность, равную 1.

В примерах с монетой и игральной костью мы без труда определяли вероятности интересующих нас событий, исходя из предположения о симметрии монеты и кубика. Понятно, что это очень частные ситуации, и в общем случае подобные соображения могут оказаться неприменимыми. Как же вычислять вероятности случайных событий? Универсального ответа на этот вопрос, конечно, не существует. Ситуации, в которых возникают случайные события, так разнообразны, механизмы, лежащие в их основе, так не похожи один на другой, что каждая сколько-нибудь содержательная задача требует специального рассмотрения. Тем не менее существует несколько достаточно общих схем определения вероятностей, в каждую из которых укладывается широкий класс разнообразных задач. Разработаны методы, позволяющие по вероятности сравнительно простых, как говорят, элементарных событий вычислять вероятность более сложных, представляющих собой сочетания простых. Некоторые из таких общих

схем и методов вычисления вероятностей мы рассмотрим в этой книжке.

В качестве иллюстраций будем приводить примеры из различных областей. Одни из них описывают реальные, встречающиеся на практике, явления, другие носят отвлеченный характер. В частности, полезными оказываются примеры с подбрасыванием монеты и игральной кости или выпиманием паугад шаров из урны. На этих простых примерах удобно выяснять общие закономерности, связанные с вероятностью. Вместе с тем при необходимости черные и белые шары в урне нетрудно заменить на дефектные и стандартные детали, на красные и белые цветы гороха, на гены A и a , на пораженные и непораженные мишени и т. п.

2. Классическая вероятность

Понятие вероятности случайного события впервые ввел Я. Бернулли в работе «Искусство предположений», вышедшей в свет в 1713 году уже после смерти автора. До этого речь шла о частоте, о числе благоприятных возможностей, о шансах и т. п. Биологи и в наше время предпочитают говорить о частотах генов, генотипов и фенотипов. Это иногда удобно из-за простоты и наглядности. Но в одном из самых авторитетных руководств по генетике (Ли Ч. Введение в популяционную генетику.— М.: Мир, 1978) можно прочесть совершенно справедливое замечание: «Обращение с частотами генов как с вероятностями значительно упрощает расчеты во многих генетических задачах».

Определение вероятности, которое сейчас называют классическим, ввел П. Лаплас. Оно настолько просто и естественно, что невольно удивляешься, почему оно не появилось много раньше. Несмотря на то что сфера применения этого определения с теоретической точки зрения не слишком широка, с его помощью удастся решить многие практические задачи.

Рассмотрим сначала два простых примера.

Пример 1. Будем подбрасывать не одну, а две монеты — монету № 1 и монету № 2. Какова вероятность одновременного выпадения двух гербов?

Перечислим возможные исходы нашего испытания:

- 1) на обеих монетах выпали номиналы;

- 2) на первой монете выпал номинал, а на второй — герб;
- 3) на первой монете выпал герб, а на второй — номинал;
- 4) на обеих монетах выпали гербы.

Из четырех исходов интересующее нас событие — выпадение двух гербов — наблюдается только в одном. Так как очевидна равновозможность всех четырех исходов, то естественно принять, что вероятность выпадения двух гербов равна $1/4$.

Пример 2. Рассмотрим урну, в которой имеется 3 белых и 5 красных шаров. Вынимаем шар наугад. Какова вероятность того, что будет вынут белый шар?

Пронумеруем шары. Пусть шары с номерами 1, 2, 3 — белые, а с номерами 4, 5, 6, 7, 8 — красные. Мы можем вынуть шар с любым номером. Следовательно, всего имеется 8 равновозможных исходов. Из них 3 исхода — появления шаров с номерами 1, 2, 3 — приведут к тому, что будет вынут белый шар. В такой ситуации естественно считать, что вероятность появления белого шара равна $3/8$, а красного — $5/8$.

Эти примеры подсказывают некоторый общий способ определения вероятности: нужно из общего числа равновозможных исходов выделить те, в которых наступает интересующее нас событие A , и это число «благоприятствующих» событию A исходов разделить на общее число исходов.

Уточним это определение, для чего введем некоторые новые понятия.

Будем говорить, что события A и B *несовместны* в данном испытании, если в результате испытания может произойти только одно из них. Например, если A — выпадение герба, а B — выпадение номинала на одной и той же монете, то A и B при однократном бросании монеты несовместны.

Если в результате испытания могут одновременно произойти оба события A и B , то будем говорить, что A и B *совместны*. Например, если A — выпадение герба на одной монете, а B — выпадение номинала на другой монете, то A и B совместны при одновременном бросании двух монет.

Будем говорить, что события A_1, A_2, \dots, A_n образуют *полную группу событий*, если они попарно несовместны и в результате испытания происходит одно из

них. Например, если A_i — выпадение i -й грани игральной кости, то A_1, A_2, \dots, A_6 — полная группа событий.

Предположим теперь, что исходом испытания может быть одно из n равновозможных событий, образующих полную группу. Пусть в результате m из этих событий происходит некоторое событие A . Говорят еще, что « m событий благоприятствует появлению события A ». Тогда по определению положим

$$P(A) = \frac{m}{n}. \quad (1)$$

Сформулированное правило и называется классическим определением вероятности.

Нетрудно проверить, что вероятность, определенная по формуле (1), равна нулю в случае невозможного события A , ($m = 0$); единице, если событие A достоверно, ($m = n$), и заключено между нулем и единицей в общем случае, ($m \leq n$). Легко убедиться также, что в примерах, приведенных в начале пункта, мы пользовались классическим определением вероятности. Добавим к этим примерам еще два.

Пример 3. Брошены две игральные кости. Какова вероятность, что сумма выпавших очков больше 10?

Так как с каждой гранью первой кости может выпасть любая грань второй кости, то всего равновозможных и попарно несовместных исходов $n = 6 \cdot 6 = 36$. Событию A благоприятствуют три исхода: $6 + 6$, $6 + 5$, $5 + 6$. Поэтому в силу классического определения

$$P(A) = \frac{3}{36} = \frac{1}{12}.$$

По пути можно найти вероятность выпадения не более 10 очков. Этому событию \bar{A} благоприятствуют все те события, которые не благоприятствуют событию A . Всего таких событий $36 - 3 = 33$. Поэтому по классическому определению

$$P(\bar{A}) = \frac{33}{36} = \frac{11}{12}.$$

Пример 4. Пусть вероятность поражения цели из одного орудия равна 0,9, а из другого — 0,8. Какова вероятность поражения цели при одновременной стрельбе из двух орудий?

Эту задачу, как и многие другие, удобно решать, перейдя от орудия и выстрелов к урнам и шарам. Итак,

пусть в первой урне имеется 10 шаров, из них 1 белый и 9 красных, а в другой урне — также 10 шаров, из них 2 белых и 8 красных. Какова вероятность, что среди вынутых двух шаров, по одному из каждой урны, по крайней мере один окажется красным? Очевидно, эта задача эквивалентна задаче с орудиями.

Перечислим все возможные исходы нашего испытания. С каждым из 10 шаров первой урны можно вынуть любой из 10 шаров второй урны. Таким образом, всего равновозможных и попарно несовместных исходов $n = 10 \cdot 10 = 100$.

Сколько исходов благоприятствует нашему событию A ? Если из первой урны вынут один из 9 красных шаров, то событие A произойдет, какой бы шар ни вынули из второй урны. Таким образом, уже имеется $m_1 = 9 \cdot 10 = 90$ благоприятствующих исходов. Если же из первой урны вынут белый шар, то событие A произойдет, когда из второй урны будет вынут один из 8 красных шаров. Таким образом, мы имеем еще $m_2 = 1 \cdot 8 = 8$ благоприятствующих исходов. Всего же таких исходов $m = m_1 + m_2 = 98$. Поэтому в силу классического определения

$$P(A) = \frac{m}{n} = \frac{98}{100} = 0,98.$$

3. Классическая вероятность в генетике

Напомним, что в клетке диплоидного организма содержится набор пар хромосом. Каждая хромосома состоит из генов. Один и тот же ген может иметь несколько вариантов. Эти варианты называются аллелями. Если аллелей два, их обозначают обычно A и a . В зависимости от того, какие аллели данного гена содержатся в хромосомной паре, могут быть три генотипа: AA , aa и Aa . Первые два из них называются гомозиготными, а третий — гетерозиготным. При скрещивании пары хромосом расщепляются на отдельные хромосомы, каждая из которых сохраняет свой ген A или a . Эти одинарные хромосомы входят в гаметы. Затем родительские гаметы объединяются в клетки потомства, образуя снова пары хромосом с тем или иным генотипом.

Часто с отдельными генами (точнее говоря, с аллелями) связаны внешне проявляемые признаки, на-

пример кареглазость или голубоглазость, курчавость или прямые волосы, леворукость или праворукость и т. д. Такое проявление признака называют фенотипом. Если данный фенотип наблюдается у генотипов AA и aA и не проявляется у генотипа aa , то ген A называют доминантным, а ген a — рецессивным. Например, у человека кареглазость, курчавость и праворукость — доминантные гены.

Рассмотрим теперь несколько простых задач из генетики. Они интересны и сами по себе как иллюстрации к классическому определению вероятности, но, кроме того, они пригодятся нам в дальнейшем.

Вероятности генов. Рассмотрим какую-нибудь диплоидную популяцию, содержащую N особей. Пусть среди них имеется D гомозигот AA , H гетерозигот Aa и R гомозигот aa . Таким образом, $D + H + R = N$. Какова вероятность встретить ген A в этой популяции? Иными словами, какова вероятность, что наугад выбранная после мейоза гамета будет содержать ген A ?

Обозначим искомую вероятность $P(A)$. Так как каждый генотип содержит по два гена, то общее количество генов в популяции, очевидно, равно $2D + 2H + 2R = 2(D + H + R) = 2N$. Так как каждый генотип AA вносит в общий набор генов два гена A и каждый генотип aA вносит один ген A , то всего генов A будет $2D + H$. Таким образом, «общий котел» генов (биологи говорят «пул генов») содержит всего $2N$ генов, из которых только $2D + H$ генов A . Отсюда по определению классической вероятности

$$P(A) = \frac{2D + H}{2N}.$$

Совершенно аналогично находится вероятность встретить ген a :

$$P(a) = \frac{2R + H}{2N}.$$

В генетике вероятности $P(A)$ и $P(a)$ обозначают соответственно p и q и называют частотами генов. Легко проверить, что $p + q = 1$.

Вероятности образования генотипов. Рассмотрим снова диплоидную популяцию с генотипами AA , aA и aa . Будем предполагать, как это часто и бывает, что возможны скрещивания между генотипами

любого вида. Такая популяция называется панмиктической.

Нас будет интересовать вероятность образования тех или иных генотипов при данном скрещивании.

Каждый генотип родительского поколения мы можем рассматривать как слово, состоящее из двух букв. Одна буква стоит на первом месте, другая — на втором. При мейозе и последующем скрещивании новый генотип может быть образован из первых букв отцовского и материнского генотипов; из первой отцовского и второй материнского; из второй отцовского и первой материнского; наконец, из вторых букв отцовского и материнского генотипов. Таким образом, всегда имеется 4 равновозможных попарно несовместных исхода, образующих полную группу. Число же исходов, благоприятствующих образованию того или иного генотипа, зависит от вида скрещивающихся генотипов.

Например, при скрещивании $Aa \times Aa$ могут быть 4 равновозможных исхода: AA , Aa , aA , aa . Если признак, контролируемый геном A , связан с полом, то генотипы Aa и aA различны. В этом случае все четыре исхода различны, и вероятность каждого из них равна $1/4$. Если же признак A с полом не связан, то генотипы Aa и aA совпадают. В этом случае генотипу AA благоприятствует один исход из четырех, генотипу $Aa = aA$ — два исхода и генотипу aa — один исход. Поэтому по классическому определению их вероятности соответственно равны $1/4$, $1/2$ и $1/4$. Заметим, что с математической точки зрения эта ситуация ничем не отличается от примера с бросанием двух монет. Роль монеты играет скрещиваемый генотип Aa , роль герба — ген A , а роль номинала — ген a .

Может случиться, что ген A доминантен. Тогда фенотипу, соответствующему этому гену, благоприятствуют три исхода: AA , aA и Aa . Поэтому вероятность появления этого фенотипа равна $3/4$, а вероятность рецессивного фенотипа aa — $1/4$.

Рассмотрим и другие варианты скрещивания. Так как любой из трех генотипов отца можно скрещивать с любым из трех генотипов матери, то всего имеется 9 вариантов скрещиваний. Чтобы описать эти варианты, удобно составить таблицу, в которой строки соответствуют генотипам отца, столбцы — генотипам матери, а на пересечении строк и столбцов указаны все 4

Таблица 1

Отец	Мать		
	АА	аА	аа
АА	АА, АА, АА, АА	Аа, АА, аА, АА	Аа, Аа, Аа, Аа
аА	аА, аА, АА, АА	аа, Аа, аА, АА	аа, Аа, аа, Аа
аа	аА, аА, аА, аА	аа, аа, аА, аА	аа, аа, аа, аа,

равновозможных исхода, полученных при данном скрещивании (первые буквы от отца и от матери, первая буква — от отца, вторая — от матери и т. д.) (табл. 1).

Пользуясь этой таблицей, легко найти вероятности тех или иных генотипов, которые могут образоваться при данном скрещивании. Например, при скрещивании $AA \times AA$ имеем, очевидно, $P(AA) = 1$, $P(aA) = P(Aa) = P(aa) = 0$. Аналогично, если генотипы aA и Aa совпадают, то например, при скрещивании $aA \times aa$ имеем $P(aa) = 1/2$, $P(Aa) = 1/2$, $P(AA) = 0$.

Теперь широко известно, что подобными скрещиваниями занимался Г. Мендель (1822—1884). Он скрестил сначала генотипы $AA \times aa$ (генотипы желтого и зеленого цвета) и получил популяцию с плодами желтого цвета, состоящую только из генотипов Aa . Повторное скрещивание $Aa \times Aa$ в этой новой популяции дало потомство, на $3/4$ желтоплодное и на $1/4$ зеленоплодное. Это соответствует вероятности фенотипа, контролируемого доминантным геном желтоплодности A (ему благоприятствуют три из четырех генотипов: AA , aA , Aa) и вероятности рецессивного генотипа aa .

Аналогичные опыты Г. Мендель проделывал с генотипами гороха, имеющими красные и белые цветы, гладкие и морщинистые плоды и т. п.

Разумеется, законы, впервые открытые Г. Менделем, справедливы не только для растений гороха, но и для всех растений, животных и человека. Так, если скрестить гомозиготный тип AA цветных мышей с гомозиготным типом aa альбиносов, то в первом поколении, состоящем из гетерозигот aA , все мыши будут цветные. А во втором поколении при скрещивании $aA \times aA$ четвертая часть потомства будет альбиносами (геп цветной окраски доминантен). Точно так же, если

у кареглазых родителей родился голубоглазый ребенок, то можно смело утверждать, что оба родителя гетерозиготны. В самом деле, ген кареглазости A доминантен, поэтому кареглазые родители могут иметь либо генотип AA , либо aA . Но скрещивания $AA \times AA$ и $AA \times aA$ не дают в потомстве генотип, соответствующий голубоглазости. Поэтому, чтобы получить такой генотип, оба родителя должны быть гетерозиготного типа aA . Тогда при скрещивании $aA \times aA$ с вероятностью $1/4$ может получиться голубоглазый генотип aa .

Рассмотрим подробнее случай, когда признак связан с полом. Это признак, ген которого содержится в половых хромосомах X или Y . При этом женские особи имеют генотип XX , а мужские — XY .

Предположим сначала, что ген, определяющий признак, находится на Y -хромосоме. Пусть этот ген, как и ранее, имеет два аллеля. Тогда соответствующие мужские генотипы можно обозначить XA и Xa . Женские особи имеют только один генотип XX . Поэтому, составляя таблицу скрещиваний, получим две строки, соответствующие генотипам отца, и только один столбец, соответствующий генотипу матери (табл. 2). В клетках этой таблицы, как и в табл. 1, указаны 4 равновероятных исхода, из которых два благоприятствуют появлению девочки (верхний ряд в каждой клетке) и два — появлению мальчика (нижний ряд в каждой клетке).

Таким образом, при любом скрещивании вероятности появления мальчика и девочки равны $1/2$. При этом мальчик будет обладать тем же признаком, что и отец (A или a), а девочка этим признаком обладать не будет. Этот совершенно очевидный вывод можно получить без всяких таблиц, но мы построили таблицу, чтобы включить и простейший случай в общую схему.

Предположим теперь, что ген, определяющий признак, содержится в X -хромосоме (гемофилия, дальтонизм и др.). Тогда этот признак может передаваться и от отца, и от матери детям обоего пола. Генотипы женских особей (XX)

Т а б л и ц а 2

Отец	Мать XX	
XA	XX AX	XX AX
Xa	XX aX	XX aX

могут быть трех видов: AA , Aa и aa , а генотипы мужских (XY) — только двух: AY и aY . Поэтому, составляя таблицу скрещиваний, получим две строки, соответствующие генотипам отца, и три столбца, соответствующих генотипам матери (табл. 3). На пересечении строк и столбцов укажем 4 равно возможных исхода, два из которых (в верхнем ряду, не содержащие Y) благоприятствуют появлению девочки и два (в нижнем ряду) — появлению мальчика. Пользуясь этой таблицей, легко найти вероятности появления мальчика или девочки с указанным признаком при данном скрещивании. Например, вероятность появления мальчика с признаком A при скрещивании $AY \times AA$ равна $P(YA) = 1/2$, а при скрещивании $AY \times Aa$ — $P(YA) = 1/4$. Точно так же вероятность появления девочки генотипа Aa при скрещивании $AY \times aa$ равна $1/2$, а при скрещивании $aY \times Aa$ — $1/4$. Если ген A доминантен, то, например, при скрещивании $AY \times Aa$ все девочки будут обладать этим признаком. Что же касается мальчиков, то лишь половина из них будут обладать доминантным признаком, а другая половина — рецессивным.

Вероятности рождения мальчиков и девочек. Вопрос о том, кто родится — мальчик или девочка — обычно остро переживается родителями и родственниками. Иногда все непременно хотят иметь сына — продолжателя рода или наследника профессии отца. В другой ситуации так же страстно желают иметь дочь — хранительницу очага. Вокруг этого вопроса издавна промышляли шарлатаны и мистификаторы, берущиеся угадать пол будущего ребенка или даже обеспечить его по заказу. Современная наука сделала в этом отношении некоторые более или менее определенные шаги. Однако рождение ребенка определенного пола в массовом масштабе — все еще случайное событие.

Т а б л и ц а 3

Отец	Мать		
	AA	Aa	aa
AY	$AA \ AA$	$AA \ Aa$	$Aa \ Aa$
	$YA \ YA$	$YA \ Ya$	$Ya \ Ya$
	$aA \ aA$	$aA \ aa$	$aa \ aa$
aY	$YA \ YA$	$YA \ Ya$	$Ya \ Ya$

Многолетние наблюдения показывают, что, как правило, мальчики рождаются несколько чаще, чем девочки. Из каждой тысячи поворожденных примерно 517 мальчиков и только 483 девочки. Отсюда следует, что вероятность рождения мальчика $P(M)$ можно считать равной 0,517, а вероятность рождения девочки $P(D)$ — равной 0,483. Однако во многих расчетах рождения мальчика и девочки можно считать равновероятными событиями и приписывать им одинаковую вероятность.

Рассмотрим несколько задач, связанных с полом ребенка. Обычно в человеческих семьях детей не очень много. Поэтому предлагаемые задачи можно решить, используя непосредственно классическое определение, вычисляя для этого общее число исходов и число благоприятствующих исходов. Каждый раз мы будем предполагать, что рождения мальчиков и девочек — равновероятные события.

Пусть в семье двое детей. Какова вероятность, что оба ребенка — мальчики? Если известно, что один мальчик, какова вероятность, что оба ребенка — мальчики? Если известно, что старший — мальчик, какова вероятность, что оба ребенка мальчики?

На первый вопрос ответить нетрудно. Имеется четыре равновероятных исхода: MM , MD , DM , DD . Исходы MD и DM различны, так как в первом из них сначала родился мальчик, а потом девочка, во втором — наоборот. Из этих четырех исходов только один MM благоприятствует нашему событию. Отсюда следует, что $P(MM) = 1/4$.

Если дополнительно известно, что один ребенок — мальчик, то событие DD исключается. Из трех равновероятных событий MM , MD , DM по-прежнему только одно MM благоприятствует желаемому исходу. Поэтому $P(MM) = 1/3$.

Если известно, что старший ребенок — мальчик, то исключаются DM и DD . В этом случае $P(MM) = 1/2$.

Подобную задачу можно поставить для семьи, имеющей более двух детей. Например, в случае трех детей имеется 8 равновероятных исходов: MMM , $ДММ$, $МДМ$, $ММД$, $ДДМ$, $ДМД$, $МДД$, $ДДД$.

Выбирая благоприятствующие исходы, можно найти вероятности двух мальчиков, не менее двух мальчиков, двух старших мальчиков, старшей девочки, среднего мальчика и т. п.

4. Некоторые формулы комбинаторики

Вернемся к предыдущей задаче. Хотя мы и заметили, что способ решения по существу не зависит от количества детей в семье, тем не менее в конкретном примере ограничились семьей лишь с тремя детьми. С ростом числа детей довольно быстро увеличивается число равновозможных исходов — т. е. число всевозможных наборов из мальчиков и девочек. Так, при десяти детях таких наборов 1024. В таком большом, хотя и конечном множестве уже не просто выделить множество благоприятствующих исходов. Подобные задачи, когда из некоторого конечного множества нужно выделить то или иное подмножество, приходится решать довольно часто не только в теории вероятностей, но и во многих других областях. Возникают вопросы о числе таких подмножеств, об их конструкции, о способах выбора составляющих подмножества элементов и т. п. Подобные вопросы решаются в специальном разделе математики, который называется комбинаторикой. В этом пункте мы приведем некоторые (самые простые) формулы комбинаторики, на которые будем опираться в дальнейшем.

Слова. Пусть имеется множество A , состоящее из конечного числа элементов: $A = \{a_1, a_2, \dots, a_m\}$. Множество A будем называть *алфавитом*, число m — *мощностью алфавита*, а элементы множества A — *буквами*.

Из элементов множества A как из букв алфавита будем составлять конечные последовательности: $a_1 a_3 a_5$; $a_2 a_7 a_2 a_1$ и т. п. Для этого на каждое место от 1 до заданного k будем помещать какой-нибудь определенный элемент множества A . Каждую такую последовательность будем называть *словом*, а число k — *длиной слова*. Например, из алфавита $A = \{0, 1\}$ мощности 2 можно составить слова единичной длины: 0 и 1, слова длиной 2: 00, 01, 10 и 11, слова длиной 3: 000, 010, 011 и т. д., слова длиной 4: 0100, 1101 и т. д. С подобными словами мы уже имели дело. Из алфавита $\{Aa\}$ мы составляли слова AA , aA и aa , из алфавита $\{MD\}$ — слова MM , MD , $ММД$ и др. Уже из приведенных примеров видно, что в общем случае длина слова может быть и больше, и меньше мощности алфавита n . Заметим, что и в русском языке есть слова, длина которых

больше мощности алфавита. Например, научное название акрихина — известного лекарства от малярии — метоксихлордизтиламинометилбутиламиноакридин содержит 44 буквы.

Пусть теперь имеется два алфавита: $A = \{a_1, a_2, \dots, a_m\}$ и $B = \{b_1, b_2, \dots, b_n\}$ мощностью m и n соответственно. Будем составлять двухбуквенные слова, помещая на первое место букву из алфавита A , а на второе — букву из алфавита B . Таким образом, получим слова a_1b_1, a_3b_7, a_mb_n и т. п. Сколько всего таких двухбуквенных слов? Так как с каждой из m букв алфавита A можно соединить любую из n букв алфавита B , то очевидно, что всего двухбуквенных слов вида a_ib_j можно образовать mn . Этим правилом пользуются с незапамятных времен, в частности, когда вычисляют площадь прямоугольника, имеющего m метров по горизонтали и n метров по вертикали.

Рассматривая три алфавита: $A = \{a_1, a_2, \dots, a_m\}$, $B = \{b_1, b_2, \dots, b_n\}$, $C = \{c_1, c_2, \dots, c_s\}$ и составляя трехбуквенные слова $a_1b_1c_1, a_1b_2c_7, a_2b_1c_3$ и т. п., нетрудно установить, что общее число таких слов равно $m \cdot n \cdot s$. Ведь с каждым из mn двухбуквенных слов a_ib_j можно соединить любую из s букв алфавита C . Обобщая, можно рассмотреть k алфавитов A_1, A_2, \dots, A_k мощностью m_1, m_2, \dots, m_k соответственно. Число Q k -буквенных слов, образованных из букв этих алфавитов, по одной из каждого, если на i -м месте ставить букву из i -го алфавита, равно произведению их мощностей:

$$Q = m_1 m_2 \dots m_k. \quad (1)$$

Это правило так часто употребляется, что имеет даже специальное название — *принцип произведения*.

Например, если A — это алфавит трех аллелей некоторого гена $A = \{a_1, a_2, a_3\}$, а B — алфавит двух аллелей некоторого другого гена $B = \{b_1, b_2\}$, то в одной неспаренной хромосоме, содержащей эти гены, может быть любой из $3 \cdot 2 = 6$ генных наборов:

$$a_1b_1, a_1b_2, a_2b_1, a_2b_2, a_3b_1, a_3b_2.$$

При объединении гамет в зиготе, т. е. при спаривании хромосом, каждый из этих шести наборов может сочетаться с таким же или с любым другим. Таким образом, может получиться $6 \cdot 6 = 36$ генотипов с дву-

мя отмеченными генами в одной парной хромосоме:

$$a_1b_1a_1b_1, a_1b_1a_1b_2, a_1b_1a_2b_1, a_1b_1a_2b_2 \quad \text{и т. д.}$$

Если признаки связаны с полом, то эти слова различны. Например, условившись писать на первом месте генный набор, т. е. гамету отца, а на втором — гамету матери, мы должны будем различать слова $a_1b_1a_2b_2$ и $a_2b_2a_1b_1$.

Аналогично можно рассмотреть хромосому с k отмеченными генами, каждый из которых имеет m_i , $i = 1, 2, \dots, k$, аллелей.

Вернемся к формуле (1). Может оказаться, что все алфавиты A_1, A_2, \dots, A_k совпадают с некоторым алфавитом A мощностью n , т. е. $A_1 = A_2 = \dots = A_k = A$. В этом случае на все k мест в слове ставятся буквы из одного и того же алфавита A . Формула (1) принимает вид

$$Q = n^k. \quad (2)$$

Таким образом, число слов длины k , образованных из алфавита мощностью n , равно n^k .

Этой формулой мы неоднократно пользовались. Так, число генотипов, т. е. слов длины 2, образованных из алфавита $\{a, A\}$ мощностью 2, равно $2^2 = 4$. Это генотипы AA, aA, Aa и aa . Число всевозможных наборов $МДД, ММД$ и т. п. из мальчиков и девочек в семье, где трое детей, — это число слов длины 3 из алфавита $\{M, D\}$ мощностью 2. Всего таких наборов $2^3 = 8$. Наконец, в только что рассмотренном примере, образуя спаренную хромосому, т. е. генотип с двумя отмеченными генами, мы образовали двухбуквенное слово, буквами в котором служили наборы генов a_1b_1, a_1b_2 и т. д. Мощность этого алфавита равна 6, поэтому всего различных генотипов с двумя отмеченными генами можно получить $6^2 = 36$.

Размещения и перестановки. Вообще говоря, как это видно из примеров, буквы в слове, составленном из одного алфавита, могут повторяться. Рассмотрим теперь важный частный случай, когда все буквы в слове различны. Слова с неповторяющимися буквами из одного алфавита называются *размещениями*. Например, из алфавита $\{0, 1\}$ можно составить только два размещения длиной 2: 01 и 10.

Число всевозможных размещений длины k , составленных из алфавита мощностью n , обозначается A_n^k

и называется кратко «число размещений из n по k ». Из формулы (1) следует, что

$$A_n^k = n(n-1)(n-2) \dots (n-k+1). \quad (3)$$

В самом деле, на первое место в размещении можно поставить любую букву из данного алфавита $A = \{a_1, a_2, \dots, a_n\}$ мощностью n . Таким образом, m_1 в формуле (1) равно n . На второе место можно поставить любую букву кроме той, которая уже стоит на первом месте. Это значит, что мощность алфавита для выбора второй буквы на единицу меньше: $m_2 = n - 1$. На третье место можно поставить любую букву кроме тех, которые уже стоят на двух первых местах. Это значит, что $m_3 = n - 2$. Продолжая таким образом, приходим, наконец, к выводу, что $m_k = n - (k - 1)$. Подставив найденные значения m_1, m_2, \dots, m_k в формулу (1), получим формулу (3).

Например, если для лечения болезни достаточно употребить любые два средства из средств 1, 2, 3 и эффект лечения зависит от последовательности применения этих средств, то можно назначить $A_3^2 = 3 \cdot 2 \cdot 1 = 6$ различных вариантов лечения. Каждому варианту соответствует некоторое двухбуквенное размещение из номеров применяемых средств:

12 21 13 31 23 32.

Формулу (3) можно записать также в виде

$$A_n^k = \frac{n!}{(n-k)!}. \quad (4)$$

В частности, когда $k = n$, получаем размещение, длина которого совпадает с мощностью алфавита. Такое размещение называется *перестановкой*. Число перестановок, которое можно получить из алфавита мощностью n , т. е. A_n^n , обозначается P_n . Из (3) при $k = n$ получаем

$$P_n = A_n^n = n(n-1) \dots 2 \cdot 1 = n!. \quad (5)$$

Например, из алфавита $A = \{1, 2, 3\}$ можно образовать $3! = 6$ различных перестановок: 123, 132, 213, 231, 312 и 321.

Сочетания. Пусть имеется некоторое конечное множество $A = \{a_1, a_2, \dots, a_n\}$ мощностью n . Любое подмножество этого множества, содержащее k элемен-

тов, называется *сочетанием из n элементов по k* . Например, из множества пяти космонавтов, каждому из которых присвоен порядковый номер 1, 2, 3, 4, 5, можно составить различные экипажи по 3 человека: {1, 2, 3}, {1, 3, 5}, {2, 3, 4}, {4, 5, 1} и т. п. Аналогично, если болезнь поддается лечению любыми двумя средствами из известных трех и эффект лечения не зависит от последовательности, в которой эти средства применяются, то можно назначить три варианта лечения: {1, 2}, {1, 3} и {2, 3}.

Число всевозможных сочетаний из n элементов по k обозначается C_n^k или $\binom{n}{k}$. Нетрудно установить, что

$$C_n^k = \frac{n!}{k!(n-k)!}. \quad (6)$$

В самом деле, для того чтобы получить всевозможные размещения длины k из алфавита мощностью n , нужно взять всевозможные сочетания из n элементов по k , а затем из каждого такого сочетания образовать $k!$ всевозможных перестановок. Таким образом,

$$A_n^k = C_n^k \cdot k!.$$

Отсюда

$$C_n^k = \frac{A_n^k}{k!}.$$

Подставив в это равенство значение A_n^k из (4), получим формулу (6).

Например, число экипажей по 3 из 5 космонавтов равно

$$C_5^3 = \frac{5!}{3!2!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 1 \cdot 2} = 10.$$

Если же экипаж составляется из четырех человек, то число различных экипажей равно

$$C_5^4 = \frac{5!}{4!1!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 1} = 5.$$

До сих пор предполагалось, что $k \geq 1$. Удобно рассмотреть также и C_n^0 — число сочетаний из k элементов, когда в сочетание берется нуль элементов. Такое сочетание — это просто пустое множество. Пустое множество единственно. Таким образом, $C_n^0 = 1$.

5. Комбинаторика и генетика

Формулы комбинаторики, с которыми мы познакомились в предыдущем пункте, помогут существенно расширить рамки применения классического определения вероятности. С их помощью не так трудно определить число всех равновозможных исходов и число исходов, благоприятствующих тому или иному событию. Рассмотрим несколько типичных задач.

Вероятность набора с одним признаком. Предположим, что в группе из 10 объектов только 8 обладают некоторым признаком. Например, 8 особей могут быть здоровыми, а две — больными; 8 особей могут обладать каким-то ценным генотипом; 8 человек из 10 могут владеть английским языком или какой-то профессией; 8 деталей могут быть кондиционными, а две — бракованными и т. п. Признак визуально, на первый взгляд обнаружить трудно. Поэтому объекты выбираются наугад. Какова вероятность, что 6 случайно выбранных объектов из 10 будут обладать указанным признаком?

Число способов, которыми 6 объектов можно выбрать из 10, как известно, равно C_{10}^6 . Таким образом, число всех равновозможных исходов $n = C_{10}^6$. Интересующее нас событие A состоит в том, что все 6 объектов выбраны из 8, обладающих указанным признаком. Число способов, которыми 6 объектов можно выбрать из 8, равно C_8^6 . Таким образом, число благоприятствующих исходов $m = C_8^6$. Отсюда по классическому определению

$$P(A) = \frac{m}{n} = \frac{C_8^6}{C_{10}^6} = \frac{8!}{6! 2!} \cdot \frac{6! 4!}{10!} = \frac{2 \cdot 3 \cdot 4}{2 \cdot 9 \cdot 10} = \frac{2}{15}.$$

Результат, как нам кажется, ошеломляющий и поучительный. Ведь группа на 80 % состоит из нужных объектов, например из здоровых людей. И требуется выбрать не всех здоровых, а только 6 из 8. Тем не менее оказывается, что лишь в 13 % случаев случайный выбор будет успешным.

Рассмотренная задача является частным случаем следующей общей ситуации. Имеется N объектов, из которых M объектов как-то отмечены (стандартные детали, здоровые особи, рабочие определенной профес-

сии, выигрышные номера лотереи и т. п.). Из N объектов наугад выбирается k объектов. Какова вероятность, что все k выбранных объектов — из числа отмеченных?

Рассуждая так же, как при решении задачи, получим $n = C_N^k$, $m = C_M^k$ и, следовательно,

$$P(A) = \frac{m}{n} = \frac{C_M^k}{C_N^k}. \quad (1)$$

Вот одна из популярных задач этого типа. Студент знает ответ на 20 вопросов из 25. Какова вероятность, что на экзамене ему попадутся 3 счастливых вопроса?

Имеем $N = 25$, $m = 20$, $k = 3$. Поэтому по формуле (1)

$$P(A) = \frac{C_{20}^3}{C_{25}^3} = \frac{20!}{3! 17!} \cdot \frac{3! 22!}{25!} = \frac{18 \cdot 19 \cdot 20}{23 \cdot 24 \cdot 25} = \frac{57}{115}.$$

Как видим, 80%-ная подготовка дает вероятность получить отличную оценку не более $1/2$.

В частном случае может быть $k = M$. Это значит, что в отбираемую группу объектов должны попасть все отмеченные.

Тогда $m = C_M^k = C_M^M = 1$, и по формуле (1) имеем

$$P(A) = \frac{1}{C_N^M}.$$

Такая ситуация с единственным шансом возникает, например, в популярной игре «Спортлото», когда нужно угадать 5 счастливых номеров из общего числа 36 номеров. В этом случае $N = 36$, $M = 5$, и следовательно,

$$P(A) = \frac{1}{C_{36}^5} = \frac{5! 31!}{36!} = \frac{2 \cdot 3 \cdot 4 \cdot 5}{32 \cdot 33 \cdot 34 \cdot 35 \cdot 36} \simeq \frac{1}{377000}.$$

К счастью, угадывать можно не весь счастливый набор, а только часть его, какие-нибудь три или четыре номера. Вероятности этих угадываний значительно больше. Мы их найдем ниже.

Другой частный случай возникает при $k = 1$. Это значит, что отбираемая группа состоит из одного объекта. Тогда

$$m = C_M^1 = M, \quad n = C_N^1 = N \quad \text{и} \quad P(A) = \frac{M}{N}.$$

Эта формула возвращает нас к простейшим примерам на классическое определение вероятности.

Вероятность набора с несколькими признаками. Рассмотрим снова группу из 10 объектов. Предположим, что 7 из них отмечены некоторым признаком. Будем говорить, что эти 7 объектов отмечены признаком A , а оставшиеся 3 объекта — признаком B . Например, A — здоровые, B — больные, A — гены A , B — гены a ; A — генотипы AA и Aa , B — генотип aa ; A — стандартные детали, B — дефектные детали и т. п. Будем по-прежнему наугад отбирать группу из 6 объектов, но потребуем, чтобы в нее вошли 4 объекта, обладающих признаком A , и 2 объекта, обладающих признаком B . Какова вероятность отбора такой группы?

Как и ранее, общее число наборов по 6 объектов из 10 равно $n = C_{10}^6$. Число наборов по 4 объекта из 7, обладающих признаком A , равно $m_1 = C_7^4$. Число наборов по 2 объекта из 3, обладающих признаком B , равно $m_2 = C_3^2$. Так как с каждым из указанных m_1 наборов с признаком A можно сочетать любой из указанных m_2 наборов с признаком B , то общее число благоприятствующих наборов $m = m_1 \cdot m_2$. Отсюда по классическому определению получаем

$$p = \frac{m}{n} = \frac{m_1 \cdot m_2}{n} = \frac{C_7^4 \cdot C_3^2}{C_{10}^6}.$$

В общем случае задача выглядит так. Имеется N объектов, из которых M объектов обладают признаком A и $N - M$ объектов обладают признаком B . Из N объектов наугад выбирается k объектов. Какова вероятность, что среди этих k объектов s объектов будут обладать признаком A и $k - s$ объектов — признаком B . Рассуждая как при решении частной задачи, получим

$$n = C_N^k, m = m_1 \cdot m_2, m_1 = C_M^s, m_2 = C_{N-M}^{k-s},$$

и, следовательно,

$$p = \frac{m}{n} = \frac{m_1 \cdot m_2}{n} = \frac{C_M^s \cdot C_{N-M}^{k-s}}{C_N^k}. \quad (2)$$

В частном случае, когда $s = k$, приходим к рассмотренной ранее задаче. В другом частном случае,

когда $k=2$, $s=1$, т. е. когда в отбираемую группу из двух объектов должны попасть по одному объекту каждого признака, получаем $m_1 = C_M^1 = M$, $m_2 = C_{N-M}^1 = N - M$, $n = C_N^2$, и, следовательно,

$$p = \frac{m_1 \cdot m_2}{n} = \frac{M(N-M)}{C_N^2}. \quad (3)$$

В качестве примера снова рассмотрим игру «Спортлото 5 из 36». Найдем вероятности угадать 4 или 3 счастливых номера. В этом случае $N=36$, $M=5$, $k=5$, а s равно либо 4, либо 3. По формуле (2) для $s=4$ находим вероятность

$$p = \frac{C_5^4 \cdot C_{31}^1}{C_{36}^5} = \frac{155}{C_{36}^5} \simeq 0,0004,$$

т. е. в 155 раз больше, чем вероятность угадать все 5 номеров. Если же $s=3$, то по той же формуле вероятность

$$p = \frac{C_5^3 \cdot C_{31}^2}{C_{36}^5} = \frac{4650}{C_{36}^5} \simeq 0,0123,$$

т. е. в 30 раз больше, чем вероятность угадать 4 номера.

Формулы (2) и (3) можно написать в уном, более симметричном, виде. Обозначим $M=N_1$, $M-N=N_2$, $s=k_1$, $k-s=k_2$. Тогда формула (2) примет вид

$$p = \frac{C_{N_1}^{k_1} \cdot C_{N_2}^{k_2}}{C_N^k}, \quad N_1 + N_2 = N, \quad k_1 + k_2 = k;$$

а формула (3) — вид

$$p = \frac{N_1 \cdot N_2}{C_N^2}, \quad N_1 + N_2 = N.$$

Теперь видно, как можно обобщить ситуацию. Пусть в группе из N объектов N_1 объектов обладают признаком A_1 , N_2 объектов — признаком A_2 и N_3 объектов — признаком A_3 , $N_1 + N_2 + N_3 = N$. Из N объектов наугад выбираются k объектов. Какова вероятность, что среди этих k объектов k_1 обладают признаком A_1 , k_2 — признаком A_2 и k_3 — признаком A_3 , $k_1 + k_2 + k_3 = k$?

Рассуждая как прежде, получим, что число всех равновозможных событий $n = C_N^k$, а число всех благоприятствующих событий (на основании принципа произведения) равно $m = m_1 \cdot m_2 \cdot m_3$, где

$$m_1 = C_{N_1}^{k_1}, \quad m_2 = C_{N_2}^{k_2}, \quad m_3 = C_{N_3}^{k_3}.$$

Отсюда по классическому определению вероятности

$$p = \frac{m}{n} = \frac{m_1 \cdot m_2 \cdot m_3}{n} = \frac{C_{N_1}^{k_1} \cdot C_{N_2}^{k_2} \cdot C_{N_3}^{k_3}}{C_N^k},$$

$$N_1 + N_2 + N_3 = N, \quad k_1 + k_2 + k_3 = k.$$

В частном случае, когда $k_1 = k_2 = k_3 = 1$, т. е. когда в отбираемую группу из трех объектов должны попасть по одному объекту каждого признака, получим

$$p = \frac{N_1 \cdot N_2 \cdot N_3}{C_N^3}, \quad N_1 + N_2 + N_3 = N.$$

Аналогично рассматриваются наборы с четырьмя, пятью и любым другим числом признаков.

Вероятности слов. Словами, составленными из букв тех или иных алфавитов, удобно кодировать разнообразные случайные явления. Например, набирая наугад трехбуквенное слово из алфавита A десяти цифр $\{0, 1, 2, \dots, 9\}$, размещенного на пульте у входной двери, мы случайно набираем слово, например 259, при котором дверь открывается. Вероятность такого единственного шанса $p = 1/n$, где n — число всевозможных трехбуквенных слов, которые можно составить из алфавита A . В зависимости от устройства когда на конструкцию слов накладываются те или иные ограничения. Например, буквы в словах не должны повторяться, т. е. слова являются размещениями из 10 по 3. В этом случае $n = A_{10}^3 = 10 \cdot 9 \cdot 8 = 720$ и, следовательно, $p = 1/720$. Дополнительное ограничение может состоять в том, что слова не должны начинаться с нуля. В этом случае $n = A_{10}^3 - 99 = 621$ и, следовательно, $p = 1/621$. Если слова могут быть любого вида, то их число $n = 10^3$ (мощность алфавита в степени длины слова) и, следовательно, $p = 0,001$.

Такая же ситуация возникает при пользовании сейфом, автоматической камерой хранения или портфелем-дипломатом с цифровым замком. Разница лишь в том, что слова состояются из четырех букв, а алфа-

вит может содержать не десять, а, например, первые пять цифр.

Подобные задачи возникают при различных сборках. Предположим, что при сборке агрегата требуется соединить в определенном порядке 5 различных деталей. Какова вероятность, что вынутые из ящика наугад одна за другой они образуют нужную последовательность?

Перенумеруем детали и предположим для конкретности, что требуемая последовательность имеет вид 12345. Это одна из возможных перестановок, образованных из алфавита {1, 2, 3, 4, 5} мощностью 5. Число таких перестановок $n = 5! = 120$. Следовательно, искомая вероятность

$$p = \frac{1}{n} = \frac{1}{120}.$$

Предположим теперь, что при сборке агрегата детали под номерами 2, 3, 4 должны идти в определенной последовательности, а первая и пятая детали взаимозаменяемы. Тогда благоприятствующими будут две последовательности 1 2 3 4 5 и 5 2 3 4 1. Поэтому по классическому определению вероятность выбора наугад нужной последовательности

$$p = \frac{m}{n} = \frac{2}{5!} = \frac{1}{60}.$$

Вместо последовательности цифр или деталей можно рассмотреть последовательности молекул, мономеров, генов, кодонов, лечебных процедур, технологических приемов и т. п.

Пусть, например, А, Б, В, Г, Д, Е — набор лечебных процедур и лекарственных препаратов, необходимых для лечения некоторой болезни. Предположим, что их можно назначать в любом порядке, за исключением последовательностей, в которых есть слог АД. Например, после препарата А нельзя сразу назначать процедуру Д, так как она разрушает препарат, еще не усвоенный организмом. Какова вероятность, что при случайном наборе процедур и лекарств будет выбрана нежелательная последовательность?

Число всех равновозможных последовательностей, очевидно, равно числу перестановок из алфавита {А, Б, В, Г, Д, Е} мощностью 6. Таким образом, $n = 6!$ Найдём число нежелательных последовательностей. Для этого слог АД будем считать одной буквой. Тогда

число всевозможных последовательностей с этим слогом равно числу перестановок из алфавита $\{B, V, Г, E, АД\}$ мощностью 5. Таким образом, $m = 5!$ Отсюда по классическому определению искомая вероятность

$$p = \frac{m}{n} = \frac{5!}{6!} = \frac{1}{6}.$$

Может быть противопоставлен не только слог АД, но и слог ДА, когда, например, препараты А и Д реагируют друг с другом до того, как их усвоил организм. В этом случае будет еще 5! нежелательных последовательностей, и, следовательно,

$$p = \frac{m}{n} = \frac{2 \cdot 5!}{6!} = \frac{1}{3}.$$

Такие же рассуждения можно применить к последовательности локусов в хромосоме или гостей за праздничным столом, если соседство некоторых из них запрещено.

Рассмотрим теперь задачу о наследовании по двум признакам. Мы уже касались этой задачи, когда рассматривали число различных генотипов с двумя отмеченными генами. Рассмотрим теперь вопрос подробнее, но предположим, что оба исследуемых гена имеют только два аллеля. Обозначим их A и a для первого гена и B и b для второго. Предположим сначала, как и ранее, что исследуемые гены содержатся в одной хромосоме. Тогда неспаренная хромосома может быть одного из четырех видов:

$$AB \quad Ab \quad aB \quad ab.$$

Эти хромосомы образуются в процессе мейоза и их вид и число зависят от вида скрещиваемых генотипов родителей, от типа расхождения хромосом при мейозе, от возможного кроссинговера, при котором хромосомы обмениваются участками перед расхождением, и от других причин. Не вдаваясь сейчас в анализ этих сложных процессов, предположим, что все четыре указанных варианта хромосом равновозможны. Как обычно, предположим, что и все возможные парные сочетания этих хромосом также равновозможны. Таких сочетаний, т. е. двухбуквенных слов из алфавита $\{AB, Ab, aB, ab\}$ мощностью 4 можно образовать $4^2 = 16$. Помещая на первом месте гениный набор отца, будем иметь $ABAB, ABAb$ и т. д.

По-видимому, правильнее было бы писать эти слова в виде схем AB, AB и т. д., отводя каждой парной хромосоме свою строчку. Однако такая двухэтажная запись, конечно, слишком громоздка. Поэтому в генетике пишут, объединяя в пары, генотипы по отдельным генам: $(AA)(BB)$, $(AA)(Bb)$ или, для краткости, без скобок $AABB$, $AABb$ и т. д. В такой записи на печатных местах стоят гены, пришедшие от отца, а на четных — от матери. Все 16 двугенных генотипов легко описать, построив таблицу, в которой строки соответствуют генным наборам отца, столбцы — генным наборам матери, а на пересечении строк и столбцов указан генотип, который получается при объединении родительских хромосом (табл. 4).

Если признаки связаны с полом, то, как мы уже говорили, все эти генотипы различны. Если связи с полом нет, то генотип определяется лишь количеством тех или иных аллелей, входящих в него. Поэтому генотипы, стоящие на симметричных местах относительно «главной диагонали», совпадают: $AAbB = AABb$, $aABb = AaBB$ и т. д. Всего различных генотипов в этом случае 9. Пользуясь таблицей и классическим определением, легко найти их вероятности:

$$P = (AABB) = P(AAbb) = P(aaBB) = P(aabb) = \frac{1}{16},$$

$$P(AAbb) = P(AaBB) = P(Aabb) = P(aaBb) = \frac{2}{16} = \frac{1}{8}.$$

$$\text{Наконец, } P(AaBb) = \frac{4}{16} = \frac{1}{4}.$$

Если ген A доминантен, то соответствующий ему фенотип встречается в 12 случаях из 16 и, следовательно, вероятность этого фенотипа $p = 12/16 = 3/4$. Аналогично для гена B .

Таблица 4

Отец	Мать			
	AB	Ab	aB	ab
AB	$AABB$	$AABb$	$AaBB$	$AaBb$
Ab	$AAbB$	$AAbb$	$AabB$	$Aabb$
aB	$aABb$	$aABb$	$aaBB$	$aaBb$
ab	$aAbB$	$aAbb$	$aabB$	$aabb$

Если доминантны и A , и B , то соответствующий фенотип по обоим генам встречается в 9 случаях из 16, а хотя бы по одному — в 11 случаях из 16. Соответствующие вероятности равны $9/16$ и $11/16$.

Из табл. 4 можно находить и другие вероятности. Например, вероятность унаследования от отца генного набора AB равна $1/4$ (первая строка таблицы); вероятность унаследования от матери гена a равна $1/2$ (третий и четвертый столбцы таблицы) и т. д.

Аналогично рассматривается задача о наследовании по двум признакам, когда гены, контролирующие эти признаки, содержатся в различных хромосомах. В этом случае надо говорить не об одной неспаренной хромосоме, а о гамете, содержащей две неспаренные хромосомы. Пусть, например, популяция состоит только из гетерозиготных особей по обоим признакам, не связанным с полом. Это означает, что все особи имеют генотип

AB

ab

Столбцы этой схемы соответствуют различным парным хромосомам, а строки — генному набору, составленному из различных непарных хромосом.

При мейозе возможны различные варианты расхождений парных хромосом к северному и южному полюсам клетки. Например, к северному полюсу могут отойти непарные хромосомы A и B . Но могут отойти также A и b , или a и B , или, наконец, a и b . В соответствии с этим могут образоваться 4 варианта гамет: AB , Ab , aB и ab . Все эти варианты равновозможны. Таким образом, мы снова имеем 4 равновозможных генных набора AB , Ab , aB и ab , из которых при слиянии родительских гамет могут образоваться 16 генотипов, указанных в табл. 4.

Пользуясь этой таблицей и классическим определением вероятности, легко найти вероятности тех или иных генотипов или фенотипов. Например, если A , a — гены курчавости и прямоволосости, B , b — гены кареглазости и голубоглазости, то поскольку A и B доминантны, имеется всего четыре фенотипа: курчавый — кареглазый; курчавый — голубоглазый; прямоволосый — кареглазый и прямоволосый — голубоглазый.

Первому из них соответствуют все генотипы табл. 4, которые содержат хотя бы одну букву *A* и хотя бы одну букву *B*. Таких генотипов, как мы уже подсчитали, 9, и, следовательно, вероятность появления курчавого кареглазого потомка равна $9/16$. Аналогично, прямоволосым кареглазым соответствуют все генотипы, которые содержат хотя бы одну букву *A* и не содержат букву *B*. Таких генотипов только три. Поэтому вероятность появления потомка указанного фенотипа равна $3/16$. Таковую же вероятность имеет прямоволосый кареглазый потомок, генотип которого содержит хотя бы одну букву *B*, но не содержит букву *A*. Наконец, прямоволосый голубоглазый потомок, которому соответствует единственный генотип *aabb*, имеет вероятность $1/16$.

Рассмотрим в заключение вероятности белков и нуклеиновых кислот. Хорошо известно, что молекула белка представляет собой длинную полимерную цепочку, составленную из 100—150 мономеров — аминокислот 20 различных видов. Свойства молекулы зависят от состава и расположения мономеров. Таким образом, молекула белка — это слово длиной 100—150, составленное из алфавита мощностью 20. Всего таких различных молекул-слов может быть не менее

$$n^k = 20^{100} = 2^{100} \cdot 10^{100} = (2^{10})^{10} \cdot 10^{100} \simeq \\ \simeq (10^3)^{10} \cdot 10^{100} = 10^{130}.$$

Аналогично молекулы ДНК представляют собой цепочки, составленные из четырех оснований нуклеиновых кислот: А, Ц, Г, Т. Хотя мощность алфавита в этом случае равна всего четырем, длина слов-цепочек обычно очень велика. Например, у всех высших растений, моллюсков, ракообразных, птиц, рептилий, рыб, амфибий, млекопитающих длина молекулы ДНК равна примерно 10^9 . Различных молекул такой длины из четырехбуквенного алфавита можно было образовать 4^{10^9} .

Число 10^{130} , тем более 4^{10^9} , исключительно велико; чтобы его представить, обычно прибегают к различным космическим сравнениям. Например, если бы удалось нашу Галактику плотно набить атомами водорода, то мы получили бы «всего» $2 \cdot 10^{82}$ атомов. Это намного меньше, чем 10^{130} .

Отсюда следует невозможность случайной сборки молекулы заданного вида, например уже собранной ранее. В самом деле, вероятность такой сборки равна вероятности одного шанса из 10^{130} или из 4^{10^9} .

У близких видов молекулы белка совпадают не полностью. Например, у японского макака и человека, как и у всех млекопитающих, α -цепь гемоглобина состоит из 141 аминокислотного остатка. Но в четырех позициях эти цепи отличаются. Могут ли такие цепи возникнуть независимо одна от другой? Поскольку в 137 позициях α -цепи макака и человека совпадают, вероятность случайного образования одной из таких цепей, если другая задана, равна $p = 1/n^4 = 1/20^{137}$. Это исчезающе малая величина.

Мы пенемного увеличим вероятность, если будем рассматривать не только близкие виды, но вообще все виды, когда-либо существовавшие и существующие на Земле. Общее количество различных белков у всех видов равно примерно 10^{12} , а нуклеиновых кислот — 10^{10} . Вероятность случайного образования белка, имеющегося в природе, равна, таким образом, $p = 10^{12}/10^{130}$, а вероятность образования молекулы ДНК — $p = 10^{10}/4^{10^9}$. Это все еще совершенно ничтожные числа.

Разумеется, молекулы белков и полинуклеотидов образуются не совсем случайно, так же, как далеко не случайна близость белков у близких видов. И найденные в рассмотренных примерах вероятности — это, конечно, не ответ, а скорее еще одно побуждение к поиску ответа на коренные вопросы о происхождении и сущности жизни, о близости и изменчивости видов. К исследованию этих важнейших проблем привлекаются громадные силы ученых различных областей. Не последнюю роль здесь играют и математические методы, в том числе и методы теории вероятностей. Но, разумеется, они много сложнее того, с чем мы успели познакомиться.

6. Геометрическая вероятность

В предыдущих разделах мы убедились, что простое, естественное, как бы само собой разумеющееся определение вероятности, которое называли классическим, оказывается вполне эффективным средством для объективной характеристики частоты наступления слу-

чайного события во многих задачах. В то же время применение этого определения ограничено некоторыми довольно жесткими условиями. Для вычисления вероятности по классической схеме нужно знать число всех равновозможных исходов, число благоприятствующих исходов и быть уверенным, что исходы попарно несовместны. Для этого нужно знать структуру исследуемых событий, как сложное событие сконструировано из простых, каковы источники благоприятствующих событий и т. п. Кроме того, во многих испытаниях число исходов так велико, что в математических моделях таких испытаний удобно считать это число бесконечным. В подобных ситуациях классическое определение неприменимо, и для вычисления вероятности прибегают к другим приемам. Один из них — геометрическая интерпретация.

Все возможные исходы изображаются точками некоторой фигуры G на плоскости или точками некоторого отрезка G на прямой, а благоприятствующие исходы — точками подмножества A множества G . По определению полагают, что вероятность события, которому благоприятствуют указанные исходы, равна отношению

$$p = \frac{|A|}{|G|}, \quad (1)$$

где прямые скобки, как обычно, обозначают меру (площадь или соответственно длину) (рис. 1).

Нельзя сказать, что геометрическое определение вероятности никак не связано с классическим. Ведь фигуру G , например, на плоскости, мы можем разбить на квадратики, столь мелкие, что с каждым случайным событием целесообразно будет связать не точку, а такой именно квадратик. Тогда число всех равновозможных исходов — это число всех квадратиков, составляющих фигуру G , т. е. площадь G , а число исходов, благоприятствующих событию A , — это число квадратиков, содержащихся в фигуре A , т. е. площадь фигуры A . Отсюда и получается формула (1).

Разумеется, число квадратиков дает лишь приближенное значение площади, и чтобы получить точное значение, нужно переходить к пределу при неограниченном измельчении квадратиков. Смысл геометрического определения и состоит в том, что не надо

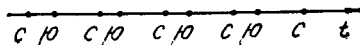
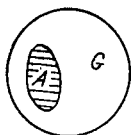


Рис. 2.

Рис. 1.

каждый раз разбивать фигуры на квадратики, вычислять их разумный размер, определять их число и переходить к пределу. Достаточно просто, пользуясь готовыми приемами, вычислить площади фигур A и G и взять отношение этих площадей. Вот несколько простых примеров.

Задача о выборе певички. Молодому человеку, живущему в центре Москвы, нравились две девушки — блондинка, живущая на севере, и брюнетка, живущая на юге. Обе девушки были одинаково привлекательны, и молодой человек никак не мог решить, какой из них сделать предложение. Наконец, в один прекрасный день он решил доверить свою судьбу случаю. Спускаясь в метро в центре, отправлялся на свидание к той девушке, чей поезд приходил первым. Через год он обнаружил, что с северной девушкой он встречался в два раза чаще, чем с южной. Этот факт юноша расценил как указующий перст судьбы и сделал предложение блондинке. Разумно ли поступил молодой человек или и на этот раз закон управлял случаем?

Нанесем на временной оси моменты прихода поездов и пометим точки прихода поездов в сторону севера буквой C , а в сторону юга — буквой $Ю$. Предположим, что интервал между поездами в обоих направлениях один и тот же, например 3 минуты. Это значит, что расстояние между двумя последовательными буквами C , как и расстояние между двумя последовательными буквами $Ю$, равно 3. Однако далее предположим, что поезд в южном направлении приходит не через полторы, а через одну минуту после поезда северного направления. Тогда буквы C и $Ю$ будут расположены так, как на рис. 2. Видно, что вероятность попасть к «северной» девушке равна вероятности попадания в промежуток $ЮC$, а к «южной» — в промежуток $СЮ$. Поскольку первый промежуток в два раза длиннее, чем второй, то и вероятность попасть к «северной» девушке в два раза больше.

Эту задачу, во всяком случае формулировку, приписывают известному советскому физическому академику Г. И. Будкеру. Говорят, что с ее помощью он проверял сообразительность своих аспирантов.

Задача из телеигры. В популярной игре «Что? Где? Когда?» стол рулетки разделен на 12 одинаковых секторов, на каждом из которых лежит конверт с вопросом. Для ответа выбирается конверт из того сектора, на который укажет стрелка рулетки. Казалось бы, все вопросы находятся в равном положении, и вероятность выбора любого из них одна и та же. Однако по ходу одной игры было замечено, что на один сектор, а именно на сектор, содержащий трудный для ответа конверт-блиц, выбор пал дважды подряд. Можно было отказаться от этого злополучного конверта, заменив ответ музыкальной паузой, что игроки и сделали. Но пауз только две, и после использования второй команда знатоков потребовала изменить расположение конвертов, уверяя ведущего, что существует большая вероятность выбора этого конверта и в третий раз. Какими соображениями руководствовались знатоки?

Дело в том, что по правилам игры после того как конверт выбран, его сектор не заполняется новым, а остается пустым. Если стрелка нового раунда укажет на пустой сектор, то для ответа выбирается конверт из ближайшего сектора по часовой стрелке. Этот пункт правил резко меняет вероятность выбора конвертов по ходу игры. В самом деле, в начале игры вероятность выбора любого конверта одна и та же и равна $1/12$. Пусть в первом раунде выбор пал на сектор № 1. Тогда перед началом второго раунда вероятность выбора любого из секторов от № 3 до № 12 равна $1/12$, а вероятность выбора сектора № 2 равна $2/12 = 1/6$. Ведь этот сектор будет выбран, если стрелка остановится либо в секторе № 2, либо в секторе № 1 (рис. 3). А так как площадь такого сдвоенного сектора в два раза больше, чем одинарного, то и вероятность его выбора в два раза больше.

Если во втором раунде будет выбран сектор № 2, то перед третьим раундом вероятность выбора сектора № 3 будет уже $3/12 = 1/4$ (этому событию будет благоприятствовать остановка стрелки в любом из секторов № 1, 2 и 3). Если после пяти раундов выбраны

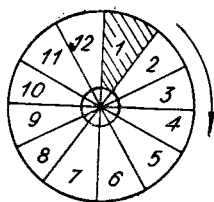


Рис. 3.

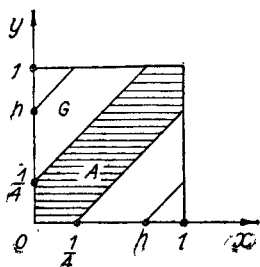


Рис. 4.

сектора с 1-го по 5-й, то в шестом раунде с вероятностью $1/2$ будет выбран сектор № 6.

В описанной реальной игре перед сектором с блит-конвертом было 4 пустых сектора, в то время как перед другими секторами был либо один пустой сектор, либо ни одного. Поэтому вероятность выбора блит-конверта ($5/12$) была намного выше, чем вероятность выбора любого другого конверта ($1/12$ или $2/12$). На это и обратили внимание знатоки.

Задача о встрече. Два приятеля договариваются встретиться между 12 и 13 часами и ждать друг друга не более 15 минут. Какова вероятность встречи?

Чтобы упростить чертеж, сдвинем условие на 12 часов и будем считать, что встреча назначена между 0 и 1. Обозначим x — момент прихода одного приятеля, а y — другого. Тогда по условию $0 \leq x \leq 1$ и $0 \leq y \leq 1$. Встреча произойдет, если $|x - y| \leq 1/4$. Таким образом, множество всех исходов G — это единичный квадрат, а множество благоприятных исходов A — это множество тех точек (x, y) из G , для которых $|x - y| \leq 1/4$ (рис. 4).

Указанное условие эквивалентно неравенствам $-1/4 \leq x - y \leq 1/4$ или $y \leq x + 1/4$, $y \geq x - 1/4$.

Построив отрезки прямых $y = x \pm 1/4$ в квадрате G , мы и найдем фигуру A благоприятных исходов. Площадь этой фигуры легко вычислить. Она равна площади G без площадей незаштрихованных треугольников:

$$|A| = 1 - 2S_{\Delta} = 1 - 2 \cdot \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{3}{4} = \frac{7}{16}.$$

Таким образом, вероятность встречи $p = |A|/|G| = 7/16$. Как видим, условия встречи установлены не очень удачно. Вероятность встречи меньше 0,5.

Как изменить условия, чтобы увеличить вероятность встречи? Очевидно, надо удлинить время ожидания. Пусть время ожидания равно h долей часа. Каким должно быть h , чтобы вероятность встречи была, например, не менее 0,99?

Рассуждая так же, как и ранее, придем к выводу, что встреча состоится, если $|x - y| \leq h$, т. е. если

$$x - h \leq y \leq x + h.$$

Площадь фигуры A , точки которой удовлетворяют написанному неравенству, равна

$$|A| = 1 - 2S_{\Delta} = 1 - 2 \cdot \frac{1}{2} (1 - h)^2 = 2h - h^2.$$

Таким образом, h нужно выбирать так, чтобы $2h - h^2 \geq 0,99$ или $1 - 2h + h^2 \leq 0,01$, т. е. $(1 - h)^2 \leq 0,01$. Отсюда находим $h \geq 0,9$.

Задачи о вероятности встречи возникают во многих областях науки и техники. Встречаться могут молекулы в активной зоне, животные конкурирующих или симбиотических видов, детали при автоматической сборке, средства транспорта, покупатели и продавцы и т. п.

Геометрические вероятности в биологии. Геометрическое определение вероятности оказывается особенно удобным в тех случаях, когда речь идет о реальных событиях на плоскости или в трехмерном пространстве. Задачи с такими событиями часто встречаются во всех областях естествознания, в том числе и в биологии.

Рассмотрим несколько примеров. Предположим, что на большом плоском участке — фигуре G — отдельными островками или даже отдельными особями расположены интересующие нас объекты, составляющие в совокупности плоскую фигуру A (рис. 5). Тогда вероятность того, что объект находится в паугад вы-

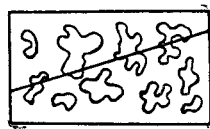


Рис. 5.

бранной точке фигуры G , равна по определению $p = |A|/|G|$. Объектами могут быть молекулы клетки, первые окончания на коже, растения, животные и т. п. Вероятность p характеризует плотность распределения объектов в фигуре G .

Площадь фигуры G обычно известна. Это — площадь национального парка, площадь мерного участка исследуемого ландшафта, площадь лабораторной стеклянной пластинки и др. Что же касается площади $|A|$, то она-то, а вместе с нею и вероятность p и являются основным предметом исследования.

Например, число объектов n в фигуре G может быть известно (число животных, запущенных в заповедник, число кровяных телец, взятых на анализ, и т. д.). Известна площадь S , занимаемая каждым объектом. Тогда очевидно, что $|A| = S \cdot n$ и, следовательно, $p = S \cdot n / |G|$. Если эта вероятность, найденная теоретически, близка к частоте встреч, которую находят экспериментально, то делают вывод о равномерном распределении объектов по всему участку. Если же между p и экспериментально найденной частотой имеется существенное различие, то это говорит о том, что особи распределены неравномерно, что имеются более и менее предпочтительные ниши, что особи предпочитают группироваться в колонии и т. п.

Часто, однако, число особей n неизвестно и его-то как раз и требуется найти. Тогда встает вопрос о методах определения площади A или вероятности p . Здесь наряду с чисто геометрическими соображениями используются приемы теории вероятностей. Вот один из распространенных методов. В фигуре наугад выбирается несколько мерных участков, например квадратов. Вычисляется площадь, занятая объектами, или число объектов в каждом таком квадрате. Берется среднее арифметическое по всем мерным квадратам и умножается на число мерных квадратов, которые могут вместиться в фигуру G . Если мерный квадрат имеет сторону с единичной длиной, то среднее арифметическое нужно умножить просто на площадь фигуры G . В этом случае квадраты называют считающими. В сельском хозяйстве вместо квадратов часто употребляют обруч, который наугад бросают, например, на поле с мелкими растениями.

Но ситуация не столь проста, как кажется на первый взгляд. Сколько нужно выбрать мерных квадратов? Сколько раз бросить обруч? Почему среднее арифметическое без искажений отразит реальную картину? Какова вероятность ошибки? Эти далеко не простые вопросы решаются методами теории вероят-

постей. С некоторыми из них мы познакомимся в дальнейшем.

Другой метод оценки площади $|A|$ основан на сечении фигуры G произвольной прямой. Если l_1, l_2, \dots, l_n — длины отрезков, по которым случайно проведенная прямая пересекает «пятна» фигуры A , расположенные в G (см. рис. 5), а L — длина этой прямой, то хорошей оценкой отношения $|A|/|G|$ может служить отношение $\Sigma l_i/L$.

Еще одна геометрическая задача описана в книге М. Кендала и П. Морана «Геометрические вероятности» (М.: Наука, 1972). Вирусная частица, представляющая собой шар, подвергается атаке антител. Антитела имеют вид тонких цилиндров и прикрепляются к поверхности вирусной частицы торцами. При этом вокруг торца прикрепленного антитела на поверхности вируса образуется зона в виде сферической шапочки, точками которой вирус не может прикрепиться к здоровой клетке. Мекает торчащее как иголка антитело. Эту зону называют тенью антитела. Предполагается, что антитела прикрепляются независимо и более или менее равномерно, причем их тени могут пересекаться. Какова вероятность, что при заданном числе антител n вся поверхность вируса будет затенена? Решение задачи оказалось очень сложным. Во всяком случае к моменту публикации книги Кендала и Морана на русском языке она была решена лишь при некоторых частных предположениях. В третьей главе мы несколько изменим постановку задачи и приведем решение в этой измененной постановке.

Закончим примером из молекулярной генетики. Известно, что перед расхождением при мейозе хромосомы обмениваются участками. Это явление называется кроссинговером. Замечено, что участки, расположенные далеко друг от друга, чаще обмениваются при кроссинговере, чем близкие, соседние участки. Найдём вероятности таких обменов, предполагая, что хромосома при кроссинговере разывается только в одном месте.

Представим парную хромосому в виде двух последовательностей генов, расположенных одна над другой:

$$A_1 \ A_2 \ \dots \ A_s \ \dots \ A_k \ \dots \ A_n$$

$$a_1 \ a_2 \ \dots \ a_s \ \dots \ a_k \ \dots \ a_n$$

Выясним возможность образования после кроссинговера непарной хромосомы, содержащей пару генов $A_s a_k$. Такая хромосома образуется в тех случаях, когда разрыв происходит сразу после гена A_s , сразу после гена A_{s+1} и т. д., наконец, сразу после гена A_{h-1} . Таким образом, среди всех $n - 1$ разрывов, которые мы считаем равновероятными, указанному типу хромосомы благоприятствует $k - 1 - s$ разрывов. Следовательно, вероятность образования такой хромосомы

$$P(A_s a_k) = \gamma \frac{k - 1 - s}{n - 1},$$

где $\gamma \leq 1$ — коэффициент, характеризующий возможность кроссинговера вообще. Переходя от генов конечной длины к точкам, что можно сделать, если хромосома достаточно длинная, а гены малы, мы вместо предыдущего равенства можем написать

$$P(A_s a_k) = \gamma \frac{d_{sk}}{L},$$

где L — длина хромосомы, а d_{sk} — расстояние между генами, стоящими на s -й и k -й позициях. Таким образом, чем дальше друг от друга расположены гены, тем больше вероятность их обмена при кроссинговере.

7. Статистическая вероятность

Еще одним приемом определения вероятности является статистический подход. От классической и геометрической схем этот метод существенно отличается тем, что вероятность вычисляется не теоретически до опыта, а экспериментально, после опыта. Статистические методы широко применяются в производстве, транспорте, экономике, медицине и естествознании в тех случаях, когда из-за сложности изучаемого явления невозможно заранее теоретически вычислить вероятность того или иного события.

Суть статистического подхода состоит в следующем. Рассмотрим достаточно длинную серию испытаний, например n_1 испытаний. Отметим случаи появления в этой серии интересующего нас события A . Пусть таких случаев было m_1 . Дробь m_1/n_1 назовем относительной частотой появления события A в данной серии. Рассмотрим новую серию испытаний примерно в тех же условиях и снова найдем относительную ча-

стоту m_2/n_2 . Аналогично для третьей, четвертой, и т. д., k -й серий испытаний найдем их относительные частоты m_3/n_3 , m_4/n_4 , ..., m_k/n_k . Может случаться, что от серии к серии относительная частота сильно колеблется. Тогда о вероятности события A ничего сказать нельзя. Однако может быть и так, что относительная частота «устойчива». Это значит, что она мало меняется от серии к серии, колеблясь около некоторого числа p . В этом случае число p и принимают за вероятность события A . Так, например, при изготовлении деталей на отлаженном станке ОТК отмечает, что в каждой партии из 1000 деталей встречается не более 4 и не менее 2 бракованных. Тогда можно утверждать, что вероятность изготовления бракованной детали равна 0,003. Если же в некоторых партиях бракованных деталей 2, 3, а в других 30—40, то о вероятности появления брака сказать ничего нельзя. Такой разницей говорит о том, что либо меняются условия работы станка (нестандартные заготовки, сбой в подаче энергии и т. п.), либо станок не отрегулирован.

Аналогично определяется вероятность поражения мишени при стрельбе из данного орудия, вероятность рождения мальчика (равная 0,518), вероятность заболевания, вероятность химической реакции и т. п. Такими же вероятностными характеристиками являются всхожесть семян, доля сердечников или аллергиков, количество опечаток на странице или несчастных случаев в течение часа, средняя рождаемость или смертность, число больных туберкулезом на 10 000 населения, количество проданных товаров за неделю и т. д.

Технические приемы определения статистической вероятности мы рассмотрим в главе, посвященной математической статистике.

8. Вероятности большие и маленькие

Даже в том небольшом списке примеров, которые мы успели рассмотреть, встретились очень маленькие вероятности ($1/10^n$, $1/376000$ и т. п.) и довольно большие (0,98, 0,99 и т. п.). Заключение о их малости или, наоборот, значительности мы делали исходя из содержания задачи, при решении которой они возникали. Это обстоятельство необходимо подчеркнуть. Вопрос о

том, какую вероятность считать малой, а какую большой, обсуждается специально еще на стадии постановки задачи. Другими словами, вероятность, маленькая в одной ситуации, оказывается недопустимо большой в другой.

В самом деле, пусть вероятность дефекта в некоторой детали равна 0,01. Велика или мала эта вероятность? Теория вероятностей на этот вопрос не может и не обязана отвечать. Все зависит от нашей договоренности, которая руководствуется конкретными практическими соображениями. Если указанная деталь используется в детском конструкторе и недорога, то можно считать вероятность дефекта, равную 0,01, малой и не менять технологии изготовления детали. В крайнем случае, можно добавить в комплект одну-две запасные детали. Но если эта же деталь, например крепёжный карабин, используется в конструкции парашюта или водолазного костюма, то вероятность ее выхода из строя, равная 0,01, должна быть признана недопустимо большой. Ведь такая вероятность означает, что примерно 10 из каждой тысячи парашютистов или водолазов могут погибнуть.

Чаще всего дело обстоит так. Исходя из практических соображений, задают необходимую вероятность сложного события, а затем уже рассчитывают другие параметры задачи, обеспечивающие заданную вероятность. Именно так мы поступили в задаче о встрече, задав вероятность ее не менее 0,99.

Аксиомы и теоремы вероятности



В различных примерах, которые рассмотрены в предыдущей главе, мы тем или иным способом определяли вероятность событий и даже выяснили кое-какие ее свойства. Иногда нам помогала интуиция, как, например, в случае статистической вероятности, а иногда — простота задачи, как в случае классической схемы.

Однако пока у нас нет общего подхода, который позволил бы ввести определение вероятности в случаях, отличных от рассмотренных. Нет критериев, опираясь на которые, мы могли бы судить, что в данной ситуации целесообразно применить вероятностные соображения.

Кроме того, часто удается, как это бывало в рассмотренных примерах, более или менее сложное событие разложить на простые. Но у нас пока нет общих правил, руководствуясь которыми мы могли бы по вероятности простых событий вычислять вероятность сложного.

Существует несколько общих схем определения вероятности. Общеизвестной в настоящее время является схема, предложенная в 20-х годах выдающимся советским математиком А. Н. Колмогоровым. В этой главе, пользуясь схемой Колмогорова (ее называют аксиоматическим подходом), обоснуем наши представления о вероятности, которые до сих пор нередко носили интуитивный характер, и докажем основные теоремы о ней. Для этого нам придется заново, теперь уже вполне строго, переопределить некоторые понятия, знакомые по предыдущей главе.

Мы подробно рассмотрим случай конечного числа исходов и сделаем краткие замечания об испытаниях, в которых исходов бесконечно много.

1. Случайные события в общей схеме

Пусть имеется множество U , состоящее из n различных элементов a_1, a_2, \dots, a_n . Это может быть множество любой природы — множество чисел, векторов, точек, слов, шаров, деталей, новорожденных, молекул, семян и т. п. Элементы a_i будем называть *элементарными событиями*, а само множество $U = \{a_1, a_2, \dots, a_n\}$ — *пространством событий*.

Из элементов множества U можно образовывать различные подмножества. Например, $A = \{a_1, a_5\}$, $B = \{a_2, a_4\}$, $C = \{a_2, a_4, a_6\}$ и т. п. Каждое такое подмножество будем называть *случайным событием*. Об элементах a_i , входящих в данное случайное событие, будем говорить, что они ему *благоприятствуют*. Например, случайному событию A благоприятствуют a_1 и a_5 , случайному событию C — a_2, a_4 и a_6 . Вместо слов «дано множество A » будем говорить «наступило событие A », или «произошло событие A ».

Если событие C наступает всякий раз, когда происходит событие B , то будем говорить, что событие B *влечет за собой* событие C . Это означает, что все элементы, благоприятствующие событию B , благоприятствуют и событию C , т. е. B является подмножеством C . В этом случае будем говорить также, что C включает B , и писать $B \subset C$.

Если $B \subset C$ и $C \subset B$, то очевидно, что B совпадает с C , так как им благоприятствуют одни и те же элементарные события.

Например, при бросании игральной кости элементарным событием a_i может быть выпадение грани, содержащей i очков. Таким образом, пространство элементарных событий U содержит 6 элементов, $U = \{1, 2, 3, 4, 5, 6\}$. Из этого пространства можно образовать различные случайные события. Например, $B = \{2, 4\}$ — выпадение либо двойки, либо четверки; $C = \{2, 4, 6\}$ — выпадение четного числа очков; $D = \{5, 6\}$ — выпадение не меньше 5 очков и др. Событию B благоприятствует любое из элементарных собы-

тий 2 или 4. Если событие B произошло, то, очевидно, произошло и событие C , т. е. C влечет за собой событие C .

Итак, несколько расплывчатые, возникшие на интуитивном уровне вероятностные термины «случайное событие», «благоприятствующее событие», «событие A влечет за собой событие B » получили четкий смысл терминов теории множеств.

Сколько различных случайных событий можно образовать из данного пространства U ? Иными словами, сколько различных подмножеств можно образовать из множества U , содержащего n элементов? На этот вопрос нетрудно ответить. Для этого выстроим элементы a_i в каком-нибудь порядке и поставим против каждого из них 1, если он входит в образуемое подмножество A и нуль — в противном случае. Тогда каждому подмножеству A будет соответствовать некоторое слово длины n , образованное из алфавита $\{0, 1\}$. Всего таких слов, как мы знаем, 2^n . Столько же и подмножеств.

Среди указанных слов будет, в частности, слово, состоящее из одних нулей. Это означает, что в образуемое подмножество A не вошло ни одно из a_i . Подмножество A пусто. Иными словами, событию A не благоприятствует ни одно из элементарных событий. Такое событие A будем называть *невозможным* и обозначать V .

Другой крайний случай — слово из одних единиц. Такому событию благоприятствует любое a_i . Каждое из них входит в образуемое множество A . Такое A будем называть *достоверным* событием. Очевидно, достоверное событие совпадает с U , поскольку, как и U , оно содержит все a_i и не содержит ничего другого.

Итак, мы имеем множество U — пространство элементарных событий, и множество Σ , содержащее 2^n различных подмножеств множества U , т. е. 2^n различных случайных событий, включая невозможное и достоверное.

В каждой конкретной ситуации пространство элементарных событий конструируется в соответствии с характером поставленной задачи. Например, если в урне имеются три шара — красный, зеленый и белый, то элементарными могут служить следующие три события: 1) вынут красный шар; 2) вынут зеленый шар; 3) вынут белый шар. Из этого пространства трех эле-

ментарных событий образуется $2^3 = 8$ случайных событий: 1) вынутый шар не является ни белым, ни красным, ни зеленым (невозможное событие); 2) вынутый шар — красный; 3) вынутый шар — зеленый; 4) вынутый шар — белый; 5) вынутый шар — не красный; 6) вынутый шар — не зеленый; 7) вынутый шар — не белый; 8) вынутый шар — либо красный, либо зеленый, либо белый (достоверное событие). Этим событиям соответствует 8 слов: 000; 100; 010; 001; 011; 101; 110; 111.

Теперь предположим, что интерес представляет не цвет шара, а факт окраски. Это может быть, например, в том случае, когда белый шар красить не нужно, а стоимость окраски в красный и зеленый цвет одинакова. Тогда в качестве элементарных целесообразно взять два события: 1) вынут белый шар, 2) вынут небелый (т. е. цветной) шар. Из этого пространства двух элементарных событий образуется $2^2 = 4$ случайных события: 1) вынутый шар не является ни белым, ни цветным (невозможное событие); 2) вынутый шар — белый; 3) вынутый шар — цветной; 4) вынутый шар — либо белый, либо цветной (достоверное событие). Этим событиям соответствуют слова: 00, 01, 10, 11.

Рассмотренный пример показывает, что применяя общую схему к конкретным ситуациям, необходимо тщательно установить, какие события рассматриваются в качестве элементарных.

2. Алгебра событий

Пусть U — пространство элементарных событий и A , B , D и т. п. — случайные события из Σ . Введем на этом множестве событий операции сложения, пересечения и вычитания.

Определение 1. Суммой событий A и B называется событие C , которое состоит в том, что произошло либо событие A , либо событие B (в частности, могут произойти и оба события A и B). Таким образом, в сумму C входят те элементарные события, которые входят либо в A , либо в B . Обозначается $C = A + B$.

Очевидно, что $A \subset A + B$ и $B \subset A + B$.

Пример 1. Пусть событие A — это попадание точки в круг, заштрихованный горизонтально, а событие

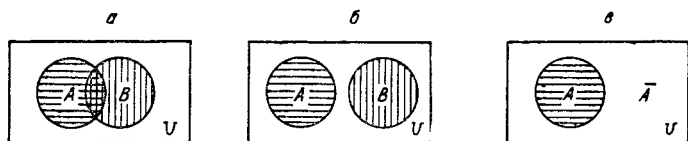


Рис. 6.

B — попадание точки в круг, заштрихованный вертикально. Тогда событие $A + B$ — это попадание точки в фигуру, покрытую хоть какой-нибудь штриховкой (рис. 6, а).

Пример 2. Пусть A — поражение мишени из первого орудия, B — из второго. Тогда $C = A + B$ — поражение мишени хотя бы из одного орудия.

Пример 3. Пусть A — выпадение герба на первой монете, B — на второй. Тогда $C = A + B$ — выпадение хотя бы одного герба при бросании двух монет.

Суммой m событий A_1, A_2, \dots, A_m называется событие C , состоящее в том, что произошло хотя бы одно из событий A_i . Обозначается $C = A_1 + A_2 + \dots + A_m$.

Определение 2. Совмещением или пересечением событий A и B называется событие C , которое состоит в том, что одновременно произошло и событие A , и событие B . Таким образом, в пересечение C входят те элементарные события, которые входят и в A , и в B . Обозначается $C = A \cdot B$ или $C = AB$.

Очевидно, $AB \subset A$ и $AB \subset B$.

В примерах 1, 2, 3 пересечение AB — это соответственно попадание точки в фигуру, покрытую двойной штриховкой, поражение мишени из обоих орудий одновременно, выпадение герба на обеих монетах.

Совмещением m событий A_1, A_2, \dots, A_m называется событие C , состоящее в том, что одновременно произошли все события A_1, A_2, \dots, A_m . Обозначается $C = A_1 \cdot A_2 \dots A_m$.

Совмещение событий называют также *произведением*.

Определение 3. События A и B называются *несовместными*, если их совмещение — невозможное событие, т. е. если $AB = \emptyset$.

Например, если A — попадание точки в круг с горизонтальной штриховкой, B — попадание в круг с вер-

тикальной штриховкой и круги не пересекаются (рис. 6, б), то A и B несовместны. Подобно этому, если A — выпадение герба, B — выпадение номинала и монета бросается один раз, то A и B несовместны.

Аналогично, m событий A_1, A_2, \dots, A_m называются несовместными, если их совмещение — невозможное событие, т. е. если

$$A_1 \cdot A_2 \dots A_m = V.$$

Определение 4. Разностью событий A и B называется событие C , которое состоит в том, что событие A произошло, а событие B не произошло. Таким образом, в разность C входят лишь те элементарные события из A , которых нет в B . Обозначается $C = A - B$.

В примере с кругами (см. рис. 6, а) разность $A - B$ — это та часть круга, покрытого горизонтальной штриховкой, которая не покрыта вертикальной штриховкой.

В частности, разность $U - A$ обозначается \bar{A} , читается «не A » и называется событием, *противоположным* событию A (рис. 6, в).

Например, при бросании одной монеты выпадение герба противоположно выпадению номинала.

Разность $\bar{V} = U - V$, очевидно, совпадает с U . Таким образом, событие, противоположное невозможному, есть достоверное событие: $\bar{V} = U$. Аналогично, событие, противоположное достоверному, есть невозможное событие: $\bar{U} = V$.

Определение 5. Будем говорить, что множества A_1, A_2, \dots, A_m образуют *полную группу событий*, если они попарно несовместны, и их сумма — достоверное событие, т. е. если $A_1 + A_2 + \dots + A_m = U$ и $A_i A_j = V$ при $i \neq j$.

Иными словами, A_1, A_2, \dots, A_m образуют полную группу, если из этих событий происходит одно и только одно.

Легко видеть, что введенные операции над событиями являются по сути дела операциями над множествами. Несущественная разница состоит лишь в том, что вместо знаков \cup и \cap употребляются $+$ и \cdot и что пустое множество (невозможное событие) обозначается V . Эти операции удовлетворяют известным из алгебры

множеств тождествам:

- 1° $A + B = B + A, \quad AB = BA;$
- 2° $(A + B) + C = A + (B + C) = A + B + C, \quad (AB)C = A(BC) = ABC;$
- 3° $(A + B)C = AC + BC;$
- 4° $A + A = A, \quad A \cdot A = A;$
- 5° $A + \bar{A} = U, \quad A \cdot \bar{A} = V;$
- 6° $A + U = U, \quad A \cdot U = A;$
- 7° $A + V = A, \quad A \cdot V = V;$
- 8° $\overline{A + B} = \bar{A} \cdot \bar{B}, \quad \overline{A \cdot B} = \bar{A} + \bar{B};$
- 9° $\overline{\bar{A}} = A.$

Большинство из этих тождеств очевидно. Другие, например 3° и 8°, также нетрудно доказать. Для этого нужно убедиться, что любой элемент множества, написанного в равенстве слева, принадлежит множеству, написанному справа, и наоборот.

Введенные операции над событиями из Σ дают снова событие из Σ . Поэтому Σ вместе с введенными операциями называется *алгеброй событий*.

Подобным образом перечисленные операции вводятся и в случае бесконечных множеств. Усложнение состоит в том, что слагаемых в сумме и сомножителей в произведении множеств может быть бесконечно много. Такие бесконечные суммы и произведения могут уже не входить в Σ .

3. Аксиомы вероятности.

Теоремы сложения

Мы ввели операции над событиями из Σ . Теперь наша задача определить вероятность. Сделать это надо, очевидно, так, чтобы по известным вероятностям событий можно было бы вычислять вероятности их сумм, пересечений и т. п.

Житейский опыт и интуиция подсказывают возможный путь. Так, например, если интересующее нас событие может произойти не только в случае A , но и в случае B , то к шансам A , видимо, надо прибавить шансы B . Это рассуждение, как говорят, «не противоречит здравому смыслу», но, конечно, не является математическим доказательством. В общем случае такое

доказательство получить невозможно. Однако у нас есть, хотя и частный, но очень важный случай, на котором мы можем проверить справедливость этого и ему подобных предположений. Это — классическое определение вероятности. Поэтому, определяя вероятность в общей схеме, будем поступать следующим образом. Выдвинутую гипотезу о свойствах вероятности сначала проверим на классической схеме и, если там она подтвердится, возьмем это предположение в качестве аксиомы для общего случая. При этом часто вместо отвлеченной классической схемы будем использовать испытания с шарами в урнах.

Начнем с самых простых свойств. В классической схеме мы установили два свойства вероятности:

- 1) вероятность события неотрицательна;
- 2) вероятность достоверного события равна 1.

Сохраним эти свойства и в общем случае.

Теперь проверим на классической схеме наше предположение о вероятности суммы.

Теорема. В классической схеме вероятность суммы двух несовместных событий равна сумме их вероятностей, т. е.

$$P(A + B) = P(A) + P(B),$$

если A и B несовместны.

В самом деле, пусть в урне n шаров, из них m_1 — красных, m_2 — зеленых, остальные белые. Пусть A — появление красного шара, B — появление зеленого шара. Очевидно, что A и B — несовместные события, и $P(A) = m_1/n$, $P(B) = m_2/n$. Найдем вероятность их суммы $C = A + B$, т. е. вероятность появления цветного (небелого) шара.

Общее число равновозможных исходов равно n . Число исходов, благоприятствующих появлению цветного шара, равно $m_1 + m_2$. Поэтому по классической схеме

$$P(A + B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B),$$

что и требовалось доказать.

Таким образом, гипотеза о вероятности суммы событий подтвердилась в классической схеме при условии, что события несовместны. Поэтому сохраним это свойство вероятности и для общего случая.

Итак, пусть имеется некоторое конечное пространство элементарных событий U и алгебра Σ случайных событий. Числовую функцию P , заданную на множестве событий Σ , назовем вероятностью, если она удовлетворяет следующим трем условиям:

A.1. $P(A) \geq 0$ для любого $A \in \Sigma$,

A.2. $P(U) = 1$,

A.3. $P(A + B) = P(A) + P(B)$, если $A \cdot B = V$.

Условия A.1 — A.3 называются *аксиомами вероятности*. Первую из них называют аксиомой неотрицательности или аксиомой монотонности, вторую — аксиомой нормируемости, наконец, третью — аксиомой аддитивности. Тройку (U, Σ, P) называют *вероятностным пространством*.

Таким образом, вероятность — это неотрицательная, нормированная, аддитивная числовая функция, определенная на событиях A из алгебры Σ пространства элементарных событий.

В случае бесконечного пространства U дополнительно требуется, чтобы бесконечные суммы и пересечения событий входили в Σ и чтобы аксиома аддитивности оставалась справедливой для бесконечной суммы.

Вернемся к конечному пространству U . Оказывается, что трех простых, как бы самоочевидных аксиом A.1 — A.3 достаточно, чтобы, опираясь на них, доказать целый ряд теорем о свойствах вероятности, не привлекая уже больше никаких интуитивных соображений. Среди этих свойств есть уже знакомые нам по классической схеме, но есть и новые.

1°. *Вероятность невозможного события равна нулю.*

В самом деле, так как U и V несовместны и $U = U + V$, то в силу A.3 имеем

$$P(U) = P(U + V) = P(U) + P(V),$$

откуда и следует требуемое.

2°. *Если B влечет за собой A , то $P(B) \leq P(A)$.*

В самом деле, если $B \subset A$, то $A = B + (A - B)$. Так как B и $A - B$ несовместны, то в силу A.3 имеем

$$P(A) = P(B) + P(A - B).$$

Отсюда и из неравенства $P(A - B) \geq 0$, справедливого в силу A.1, следует требуемое.

Таким образом, более обширному множеству элементарных событий соответствует большая вероятность.

Поэтому это свойство называют монотонностью вероятности.

3°. Для любого A справедливо неравенство $0 \leq P(A) \leq 1$.

В самом деле, левое неравенство — это аксиома А.1, а правое следует из включения $A \subset U$, предыдущего свойства и аксиомы А.2.

4°. Если A_1, A_2, \dots, A_m попарно несовместны, то $P(A_1 + A_2 + \dots + A_m) = P(A_1) + P(A_2) + \dots + P(A_m)$. (1)

Рассмотрим для простоты случай $m = 3$. Поскольку A_1, A_2, A_3 попарно несовместны, события $A_1 + A_2$ и A_3 также несовместны. Поэтому в силу А.3 можем написать

$$\begin{aligned} P(A_1 + A_2 + A_3) &= P(A_1 + A_2) + P(A_3) = \\ &= P(A_1) + P(A_2) + P(A_3). \end{aligned}$$

Для произвольного m доказательство аналогично.

5°. Если события A_1, A_2, \dots, A_m образуют полную группу, то

$$P(A_1) + P(A_2) + \dots + P(A_m) = 1. \quad (2)$$

В самом деле, в силу полноты группы события A_i попарно несовместны и $A_1 + A_2 + \dots + A_m = U$. Отсюда и из предыдущего свойства следует требуемое.

6°. Сумма вероятностей взаимно противоположных событий равна 1:

$$P(A) + P(\bar{A}) = 1. \quad (3)$$

Это свойство следует из предыдущего, так как события A и \bar{A} образуют полную группу.

7°. Вероятность суммы событий равна сумме их вероятностей без вероятности совмещения:

$$P(A + B) = P(A) + P(B) - P(AB). \quad (4)$$

Прежде чем доказывать эту теорему, заметим, что ее формулировка в отличие от А.3 не предполагает несовместности событий A и B . Если же они несовместны, то их совмещение — невозможное событие, его вероятность равна нулю и формула (4) превращается в А.3.

Переходим к доказательству. Так как $A + B = A + (B - AB)$ и события A и $(B - AB)$ несовместны, то

$$P(A + B) = P(A) + P(B - AB). \quad (5)$$

Так как $B = (B - AB) + AB$ и события $(B - AB)$ и AB несовместны, то

$$P(B) = P(B - AB) + P(AB).$$

Выразив отсюда $P(B - AB)$ и подставив в (5), получим требуемое.

4. Независимость событий.

Условная вероятность

Рассмотрим сначала пример из классической схемы. Пусть в урне 3 белых шара и 2 черных. Пусть A — появление белого шара при первом извлечении, а B — появление белого шара при втором извлечении. Предположим сначала, что первый вынутый шар возвращается в урну. Тогда, очевидно, вероятность B не зависит от того, какой шар был вынут при первом извлечении. В таких случаях говорят, что *событие B не зависит от события A* .

Предположим, теперь, что вынутый первым шар не возвращается в урну. Тогда ситуация меняется. Если был вынут белый шар, т. е. событие A произошло, то вероятность появления белого шара при втором извлечении равна 0,5. Если же был вынут черный шар, т. е. событие A не произошло, то вероятность появления белого шара при втором извлечении равна 0,75. Таким образом, вероятность события B принимает разные значения в зависимости от того, произошло или не произошло событие A . В таких случаях говорят, что *событие B зависит от события A* . Вероятность события B при условии, что произошло событие A , называют *условной вероятностью* и обозначают $P(B/A)$ или $P_A(B)$.

В рамках классической схемы нетрудно установить важные формулы для вероятностей зависимых и независимых событий.

Теорема. *В классической схеме вероятность совмещения двух независимых событий равна произведению их вероятностей, т. е.*

$$P(AB) = P(A) \cdot P(B), \quad (6)$$

если A и B независимы.

В самом деле, пусть в одной урне n_1 шаров, из них m_1 белых, остальные черные; в другой урне n_2 шаров,

из них m_2 белых, остальные черные. Пусть событие A — появление белого шара из первой урны, а событие B — появление белого шара из второй. Тогда очевидно, что события A и B независимы и $P(A) = m_1/n_1$, а $P(B) = m_2/n_2$. Найдем вероятность совмещения AB , т. е. вероятность появления белого шара и из первой урны, и из второй.

Так как с каждым шаром первой урны может быть вынут любой шар второй урны, то общее число равно-возможных исходов равно $n_1 \cdot n_2$. Так как с каждым белым шаром первой урны может быть вынут любой белый шар второй урны, то общее число благоприятствующих исходов равно $m_1 \cdot m_2$. Поэтому по классической схеме

$$P(AB) = \frac{m_1 \cdot m_2}{n_1 \cdot n_2} = \frac{m_1}{n_1} \cdot \frac{m_2}{n_2} = P(A) \cdot P(B),$$

что и требовалось доказать.

Теорема. В классической схеме вероятность совмещения двух зависимых событий равна произведению вероятности одного на условную вероятность другого при условии, что первое произошло, т. е.

$$P(AB) = P(A) \cdot P(B/A). \quad (7)$$

В самом деле, пусть в урне n шаров, из них m белых, остальные черные. Пусть A — появление белого шара при первом извлечении, а B — появление белого шара при втором извлечении, причем вынутый первым шар в урну не возвращается. Тогда очевидно, что B зависит от A ; $P(A) = m/n$, $P(B/A) = (m-1)/(n-1)$. Найдем вероятность совмещения AB , т. е. вероятность появления белого шара и при первом, и при втором извлечении.

Так как с каждым из n шаров при первом извлечении можно вынуть любой из $n-1$ оставшихся шаров при втором извлечении, то общее число равно-возможных исходов равно $n(n-1)$. Так как с каждым из m белых шаров при первом извлечении можно вынуть любой из $m-1$ оставшихся белых шаров при втором извлечении, то общее число благоприятствующих исходов равно $m(m-1)$. Поэтому по классической схеме

$$P(AB) = \frac{m(m-1)}{n(n-1)} = \frac{m}{n} \cdot \frac{m-1}{n-1} = P(A) \cdot P(B/A),$$

что и требовалось доказать.

Итак, мы доказали, что в классической схеме для независимых событий A и B выполняется равенство (6), а для зависимых — равенство (7). Эти равенства называют *формулами умножения вероятностей*. Переходя к общему случаю, возьмем эти равенства за определения. Таким образом, в общей схеме события A и B будем называть независимыми, если

$$P(AB) = P(A) \cdot P(B), \quad (8)$$

а условную вероятность $P(B/A)$ определим равенством

$$P(B/A) = \frac{P(AB)}{P(A)}, \quad (9)$$

если $P(A) \neq 0$.

Введя вероятность по формуле (9), мы должны проверить, что для нее выполняются аксиомы А.1 — А.3. Это нетрудно сделать. Зафиксируем A и рассмотрим вероятность $P_A(B)$. Прежде всего отметим, что $P_A(B) \geq 0$. Далее при $B = U$ имеем

$$P_A(U) = \frac{P(AU)}{P(A)} = \frac{P(A)}{P(A)} = 1.$$

Наконец, если события B_1 и B_2 несовместны, то AB_1 и AB_2 также несовместны. Поэтому $P(AB_1 + AB_2) = P(AB_1) + P(AB_2)$. Пользуясь этим, можем написать

$$\begin{aligned} P_A(B_1 + B_2) &= \frac{P(A(B_1 + B_2))}{P(A)} = \frac{P(AB_1 + AB_2)}{P(A)} = \\ &= \frac{P(AB_1) + P(AB_2)}{P(A)} = \frac{P(AB_1)}{P(A)} + \frac{P(AB_2)}{P(A)} = \\ &= P_A(B_1) + P_A(B_2). \end{aligned}$$

Таким образом, для вероятности $P_A(B)$ выполняются все три аксиомы.

Из формулы (9) следует, очевидно, формула умножения

$$P(AB) = P(A) \cdot P(B/A), \quad (10)$$

установленная нами в классической схеме.

Формула (8) симметрична относительно A и B . Пользуясь этим, определение независимости событий можно обобщить на случай более чем двух событий. События A_1, A_2, \dots, A_m называются *независимыми в совокупности*, если для любого набора этих событий

вероятность произведения равна произведению вероятностей:

$$P(A_k A_r \dots A_s) = P(A_k) \cdot P(A_r) \dots P(A_s). \quad (11)$$

Например, $P(A_1 A_3 A_4) = P(A_1) P(A_3) P(A_4)$;
 $P(A_2 A_5 A_8 A_{10}) = P(A_2) P(A_5) P(A_8) P(A_{10})$ и т. д. В частности,

$$P(A_1 A_2 \dots A_m) = P(A_1) \cdot P(A_2) \dots P(A_m). \quad (12)$$

Рассмотрим некоторые свойства вероятности независимых событий и условной вероятности.

1°. Если A и B независимы, то

$$P(A + B) = P(A) + P(B) - P(A) \cdot P(B). \quad (13)$$

Это равенство следует из формулы (4) и условия независимости (8).

2°. Если A и B независимы, то \bar{A} и \bar{B} также независимы.

В самом деле, из тождества $\bar{A} \cdot \bar{B} = \overline{A + B}$ и формулы (3) следует, что

$$\begin{aligned} P(\bar{A} \cdot \bar{B}) &= P(\overline{A + B}) = 1 - P(A + B) = 1 - P(A) - \\ &- P(B) + P(A) \cdot P(B) = (1 - P(A)) (1 - P(B)) = \\ &= P(\bar{A}) \cdot P(\bar{B}). \end{aligned}$$

Свойство 2° имеет место и в случае независимых в совокупности m событий.

3°. Если A_1, A_2, \dots, A_m независимы в совокупности, то вероятность наступления хотя бы одного из этих событий (т. е. вероятность суммы) вычисляется по формуле

$$P(A_1 + A_2 + \dots + A_m) = 1 - P(\bar{A}_1) \cdot P(\bar{A}_2) \dots P(\bar{A}_m). \quad (14)$$

В самом деле, событие $\bar{A}_1 \cdot \bar{A}_2 \dots \bar{A}_m$ состоит в том, что не произошло ни одно из событий A_i . Оно противоположно событию, состоящему в том, что произошло хотя бы одно из событий A_i , т. е. сумме событий $A_1 + A_2 + \dots + A_m$. Поэтому в силу (3)

$$P(A_1 + A_2 + \dots + A_m) = 1 - P(\bar{A}_1 \cdot \bar{A}_2 \dots \bar{A}_m).$$

Отсюда, пользуясь тем, что для независимых в совокупности событий $P(\bar{A}_1 \cdot \bar{A}_2 \dots \bar{A}_m) = P(\bar{A}_1) \cdot P(\bar{A}_2) \dots \dots P(\bar{A}_m)$, получим требуемое.

4°. Если A и B независимы и их вероятности отличны от нуля, то

$$P(B/A) = P(B), \quad P(A/B) = P(A). \quad (15)$$

В самом деле,

$$P(B/A) = \frac{P(AB)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B).$$

Аналогично доказывается и второе равенство (15). Таким образом, в случае независимых событий условная вероятность совпадает с безусловной, что вполне согласуется с содержанием понятия независимости событий.

5°. Для любых A и B , вероятности которых отличны от нуля, справедливо равенство

$$P(A) \cdot P(B/A) = P(B) \cdot P(A/B). \quad (16)$$

В самом деле, из равенства (10)

$$P(A) \cdot P(B/A) = P(AB) = P(BA) = P(B) \cdot P(A/B).$$

6°. Если событие B может произойти только при наступлении одного из попарно несовместных событий A_1, A_2, \dots, A_m , то вероятность события B вычисляется по формуле

$$P(B) = P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2) + \dots + P(A_m) \cdot P(B/A_m). \quad (17)$$

В самом деле, по условию $B \subset (A_1 + A_2 + \dots + A_m)$ (рис. 7). Поэтому

$$B = B(A_1 + A_2 + \dots + A_m) = A_1B + A_2B + \dots + A_mB.$$

Слагаемые в правой части попарно несовместны, поэтому по формуле сложения (1)

$$\begin{aligned} P(B) &= P(A_1B + A_2B + \dots + A_mB) = \\ &= P(A_1B) + P(A_2B) + \dots + P(A_mB). \end{aligned}$$

Применив к каждому слагаемому формулу умножения (10), получим требуемое.

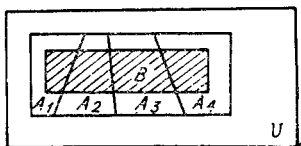


Рис. 7.

Равенство (17) называют формулой полной вероятности. Оно хорошо иллюстрируется рис. 7.

7°. Если событие B может произойти только при наступлении одного из по-

парно несовместных событий A_1, A_2, \dots, A_m , то условная вероятность $P(A_i/B)$ вычисляется по формуле

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}, i = 1, 2, \dots, m, \quad (18)$$

где $P(B)$ определяется равенством (17).

Формулы (18) следуют из равенства (16), если его записать для A_i и B . Они называются формулами Байеса по имени английского математика, опубликовавшего их в 1764 году. Формулы Байеса доставляют дополнительную информацию о событиях A_i , после того как событие B произошло. События A_i называют в данном случае *гипотезами*.

Итак, мы ввели понятие вероятности с помощью аксиом А.1 — А.3, понятие независимости событий с помощью формул (8) и (22) и понятие условной вероятности с помощью формулы (9). Опираясь на эти аксиомы и определения, мы доказали теоремы о свойствах вероятности. Всюду в дальнейшем, рассматривая какие бы то ни было вероятности, будем предполагать, что все аксиомы и, следовательно, все свойства вероятности имеют место. Приведем еще раз сводку важнейших из них в удобной для приложений форме.

Аксиомы:

1. $P(A) \geq 0$.
2. $P(U) = 1$.
3. $P(A+B) = P(A) + P(B)$, если $A \cdot B = V$.

Определения (формулы умножения):

1. События A и B называются независимыми, если $P(AB) = P(A)P(B)$.

2. События A_1, A_2, \dots, A_m называются независимыми в совокупности, если

$$P(A_k A_r \dots A_s) = P(A_k) \cdot P(A_r) \dots P(A_s)$$

для любого набора индексов k, r, \dots, s . В частности,

$$3. P(A_1 A_2 \dots A_m) = P(A_1) \cdot P(A_2) \dots P(A_m).$$

4. Условной вероятностью события B при условии, что A произошло, называется вероятность

$$P(B/A) = \frac{P(AB)}{P(A)}, P(A) \neq 0.$$

Отсюда

$$5. P(AB) = P(A)P(B/A).$$

Теоремы (формулы сложения):

1. Если A_1, A_2, \dots, A_m попарно несовместны, то
$$P(A_1 + A_2 + \dots + A_m) = P(A_1) + P(A_2) + \dots + P(A_m).$$

2. Если A и B совместны, то

$$P(A + B) = P(A) + P(B) - P(AB).$$

В частности, для независимых совместных A и B .

3. $P(A + B) = P(A) + P(B) - P(A) \cdot P(B).$

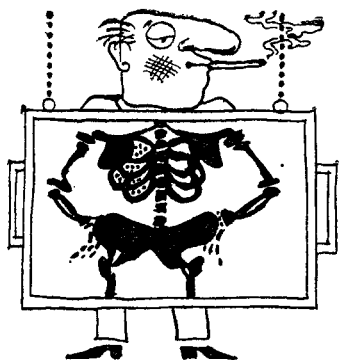
4. Если A_1, A_2, \dots, A_m независимы в совокупности, то вероятность наступления хотя бы одного события вычисляется по формуле

$$P(A_1 + A_2 + \dots + A_m) = 1 - P(\bar{A}_1) \cdot P(\bar{A}_2) \cdot \dots \cdot P(\bar{A}_m).$$

5. Полная вероятность вычисляется по формуле

$$P(B) = P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2) + \dots \\ \dots + P(A_m) \cdot P(B/A_m).$$

В следующей главе мы рассмотрим ряд задач из биологии и химии, решение которых опирается на перечисленные формулы. Некоторые из этих задач непосредственно связаны с реальными явлениями. В других, чтобы подчеркнуть их универсальность, говорится о шарах, деталях и т. п. При необходимости эти отвлекаемые объекты легко заменить молекулами, веществами, зёрнами, особями, генотипами и т. п.



Случайные события в биологии

1. Вероятность суммы и пересечения признаков

Случайным событием в этом типе задач является наличие того или иного признака. Признак может проявиться в результате определенного вида скрещивания и, следовательно, образования определенного генотипа, в результате мутации, доминирования и т. п.

Предположим, что вследствие двух мутагенных воздействий на популяцию родителей, например под действием химических препаратов и облучения, в популяции потомства возникают два типа мутаций. Будем для определенности говорить о популяции плодовых мушек и о мутациях крыльев (K) и глаз (G).

Предположим сначала, что определенное воздействие вызывает определенную мутацию. Например, облучение вызывает только мутацию крыльев, а химические препараты — только мутацию глаз. Это означает, что мутации G и K возникают независимо одна от другой.

Пусть после облучения мушек 30 % особей в популяции потомства получают мутацию крыльев, а после химического воздействия 20 % получают мутацию глаз. Исходя из этих данных можно ставить и решать различные задачи. Например,

1. Какова вероятность, что наугад выбранная мушка из потомства имеет обе мутации?
2. Какова вероятность, что наугад выбранная мушка имеет мутацию глаз, но не имеет мутацию крыльев?
3. Какова вероятность, что наугад выбранная мушка имеет хотя бы одну мутацию?

По условию $P(K) = 0,3$, $P(G) = 0,2$. События K и G независимы. Поэтому $P(KG) = P(K) \cdot P(G) = 0,3 \cdot 0,2 = 0,06$. Тем самым мы ответили на первый вопрос.

Далее, события G и K также независимы. Поэтому $P(G\bar{K}) = P(G) \cdot P(\bar{K}) = P(G) \cdot (1 - P(K)) = 0,2 \times (1 - 0,3) = 0,2 \cdot 0,7 = 0,14$, и мы получили ответ на второй вопрос. Наконец, по третьей формуле сложения $P(G + K) = P(G) + P(K) - P(GK) = 0,3 + 0,2 - 0,3 \cdot 0,2 = 0,44$, и это ответ на третий вопрос.

Изменим теперь условие задачи. Предположим, что оба вида воздействия могут вызвать обе мутации, т. е. события G и K зависимы. В этом случае мы не можем вычислить $P(GK)$ по вероятностям $P(G)$ и $P(K)$ и, чтобы ответить на два последних вопроса, надо знать заранее ответ на первый вопрос.

Пусть по-прежнему $P(K) = 0,3$, $P(G) = 0,2$ и дополнительно известно, что половина мушек, имеющих мутацию глаз, имеют и мутацию крыльев. Это означает, что $P(GK) = 0,2 \cdot 0,5 = 0,1$.

Пользуясь этим, по второй формуле сложения сразу получаем ответ на третий вопрос:

$$P(G + K) = P(G) + P(K) - P(GK) = 0,3 + 0,2 - 0,1 = 0,4.$$

Чтобы ответить на второй вопрос, заметим, что

$$G = G(K + \bar{K}) = GK + G\bar{K}.$$

Так как события GK и $G\bar{K}$ несовместны, то

$$P(G) = P(GK + G\bar{K}) = P(GK) + P(G\bar{K}).$$

Отсюда

$$P(G\bar{K}) = P(G) - P(GK) = 0,2 - 0,1 = 0,1.$$

Вместо мутаций можно было бы рассмотреть две болезни от двух причин, побочные действия двух лекарств, дефекты изделий от разного вида обработок и т. п.

Прежде чем рассмотреть еще одну биологическую задачу с зависимыми событиями, обратимся к более простой задаче с игральной костью. Пусть, как обычно, грани игральной кости пронумерованы, например, по количеству очков на них. Предположим, кроме того, что первая грань покрашена в белый цвет, вторая и третья — в красный, четвертая и пятая — в желтый, а шестая — наполовину в красный, наполовину — в желтый.

Пусть A — событие, состоящее в том, что при бросании кости выпал красный цвет, а B — событие, со-

стоящее в том, что выпал желтый цвет. Найдем вероятность суммы $A + B$, т. е. вероятность того, что выпал либо красный, либо желтый цвет. Так как красный цвет имеется на трех гранях из шести, то $P(A) = 3/6 = 1/2$. Аналогично $P(B) = 3/6 = 1/2$. Совмещение AB имеется только на одной грани. Поэтому $P(AB) = 1/6$. Тогда по второй формуле сложения получаем

$$P(A + B) = P(A) + P(B) - P(AB) = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6}.$$

Из содержания задачи ясно, что события A и B не являются независимыми. Это подтверждается и неравенством

$$P(AB) = \frac{1}{6} \neq \frac{1}{4} = P(A) \cdot P(B).$$

Рассмотрим теперь аналогичный пример из генетики. Предположим, что некоторый ген имеет три аллеля: A_1, A_2, A_3 . Признаки, контролируемые генами A_1 и A_3 , например красный и желтый цвета, проявляются как в гомозиготах, так и в гетерозиготах, а признак, контролируемый геном A_2 , проявляется только в гомозиготе. Какова вероятность образования фенотипа, содержащего либо желтый, либо красный цвет?

Составим решетку Пеннета

	A_1	A_2	A_3
A_1	A_1A_1	A_1A_2	A_1A_3
A_2	A_2A_1	A_2A_2	A_2A_3
A_3	A_3A_1	A_3A_2	A_3A_3

Пусть A — появление фенотипа с красным цветом, а B — появление фенотипа с желтым цветом. Событию A благоприятствуют все генотипы, содержащие A_1 ; их 5 из 9 возможных (генотипы первого столбца и первой строки). Поэтому $P(A) = 5/9$. Событию B благоприятствуют все генотипы, содержащие A_3 ; их также 5 из 9 возможных (генотипы третьей строки и третьего столбца). Поэтому $P(B) = 5/9$. Совмещению AB благоприятствуют только два генотипа A_1A_3 и A_3A_1 . Поэтому $P(AB) = 2/9$. Подставив эти вероятности в фор-

мулу сложения, получим

$$P(A + B) = P(A) + P(B) - P(AB) = \frac{5}{9} + \frac{5}{9} - \frac{2}{9} = \frac{8}{9}.$$

События A и B зависимы, что подтверждается неравенством

$$P(AB) = \frac{2}{9} \neq \frac{25}{81} = P(A) \cdot P(B).$$

Закончим простой, но актуальной задачей о вреде курения. В группе обследуемых 1000 человек. Из них 600 курящих и 400 некурящих. Среди курящих 240 человек имеют те или иные заболевания легких. Среди некурящих легочных больных 120 человек. Являются ли курение и заболевание легких независимыми событиями?

Пусть A означает, что обследуемый курит, а B — что обследуемый страдает заболеванием легких. Тогда

$$P(AB) = \frac{240}{1000} = 0,24; \quad P(A) = \frac{600}{1000} = 0,6;$$

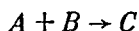
$$P(B) = \frac{240 + 120}{1000} = 0,36.$$

Так как $P(AB) = 0,24 \neq 0,6 \cdot 0,36 = P(A)P(B)$, то события A и B зависимы.

2. Теория встреч

Модели многих химических и биологических процессов опираются на так называемую теорию встреч. Согласно этой теории скорость процесса определяется числом встреч в единицу времени тех или иных объектов, участвующих в процессе (молекул, особей, гамет), и число встреч пропорционально произведению концентраций, численностей или частот этих объектов. Обоснуем теорию встреч с точки зрения теории вероятностей. Начнем с задачи из химии.

Встречи молекул. В силу закона действующих масс, открытого опытным путем Гульдбергом и Вааге, скорость реакции



пропорциональна произведению концентраций x и y веществ A и B соответственно:

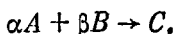
$$W = kxy.$$

К этому же выводу приводят простые вероятностные соображения.

В самом деле, предположим для определенности, что A и B — газы. Обозначим m_1 и m_2 — объемы, занимаемые соответственно A и B в реакторе, и пусть $n = m_1 + m_2$ — объем реакционной смеси. Тогда $x = m_1/n$ и $y = m_2/n$ — объемные концентрации веществ A и B . Предположив, что A и B идеально перемешаны, мы можем трактовать x и y как вероятности нахождения молекул A и B соответственно в данной точке реакционного пространства. Чтобы реакция произошла, нужно, чтобы молекулы A и B встретились в одной точке. Вероятность такой встречи по первой формуле умножения вероятностей равна произведению вероятностей. Таким образом, вероятность встречи молекул A и B равна произведению xy . Что же касается скорости реакции, то она, естественно, пропорциональна вероятности таких встреч:

$$W = kxy.$$

Аналогично рассматривается реакция



Для того чтобы такая реакция произошла, нужно, чтобы α молекул вещества A встретились с β молекулами вещества B . Вероятность такой встречи по третьей формуле умножения вероятностей равна произведению

$$\underbrace{x \cdot x \dots x}_{\alpha \text{ раз}} \cdot \underbrace{y \cdot y \dots y}_{\beta \text{ раз}} = x^\alpha y^\beta.$$

Скорость реакции, как и ранее, пропорциональна вероятности встреч:

$$W = kx^\alpha y^\beta,$$

Исходя из объемных концентраций, мы трактовали x и y как геометрические вероятности. Пользуясь законом Авогадро, мы могли бы вместо объемов рассматривать числа молекул m_1 и m_2 веществ A и B . В этом случае концентрации $x = m_1/n$, $y = m_2/n$, $m_1 + m_2 = n$, естественно, трактуются как вероятности классической схемы.

Встречи гамет. Перейдем теперь к биологическим задачам. Подобно встречам молекул в химической реакции можно было бы говорить о встречах

особей разных видов, например в системе хищник — жертва, или особях разного пола в процессе размножения, или особях, конкурирующих за одну пищу, и т. п. Рассмотрим более подробно встречи генов в зиготах.

Пусть родительская популяция содержит N особей, из которых D особей — гомозиготы AA , H особей — гетерозиготы Aa и R особей — гомозиготы aa . Тогда, как мы уже отмечали, частота гена A в популяции равна $p = P(A) = (2D + H)/2N$, а частота гена a равна $q = P(a) = (2R + H)/2N$, $p + q = 1$.

Частоты p и q мы можем трактовать как вероятности того, что наугад выбранная гамета содержит ген A или соответственно ген a . Гаметы с генами A и a , объединяясь при встречах в зиготе, порождают генотипы нового поколения. Обозначим $P(AA)$, $P(aA)$, $P(Aa)$ и $P(aa)$ вероятности образования соответствующих генотипов. (На первом месте, как обычно, помещаем ген, пришедший от отца.) Чтобы образовалась зигота AA , нужно, чтобы каждая из наугад взятых гамет содержала ген A . Так как гаметы «берутся» независимо одна от другой, вероятность такой встречи гамет в зиготе равна произведению их вероятностей (частот). Таким образом,

$$P(AA) = P(A) \cdot P(A) = p \cdot p = p^2.$$

Совершенно аналогично получаем

$$P(aA) = P(a)P(A) = qp, \quad P(Aa) = P(A)P(a) = pq,$$

$$P(aa) = P(a)P(a) = q^2.$$

Если генотипы aA и Aa неразличимы, то, поскольку образования этих генотипов — несовместные события, по аксиоме аддитивности получаем

$$P(aA + Aa) = P(aA) + P(Aa) = qp + pq = 2pq.$$

В дальнейшем в случае совпадения генотипов aA и Aa вместо $P(aA + Aa)$ будем писать просто $P(Aa)$ и, следовательно, $P(Aa) = 2pq$.

Если аллель A доминантный, то имеется два фенотипа aa и $AA + Aa$ с вероятностями $P(aa) = q^2$ и $P(AA + Aa) = P(AA) + P(Aa) = p^2 + 2pq = p(p + 2q) = p(p + q + q) = p(1 + q) = (1 - q)(1 + q) = 1 - q^2$.

Рассмотрим теперь образование генотипов с двумя признаками. Такие генотипы образуются при встрече

в зиготе гамет с двумя отмеченными генами. Будем считать для определенности, что отмеченные гены имеют по два аллеля A, a и B, b и находятся в различных хромосомах. Возможные варианты образования генотипов приведены в табл. 4. Каждое такое образование мы снова можем рассматривать как результат встречи двух гамет: гаметы AB отца с гаметой AB матери; гаметы Ab отца с гаметой aB матери и т. п. Поскольку гаметы для встречи берутся случайно и независимо одна от другой, вероятность такой встречи равна произведению вероятностей встречающихся гамет:

$$P(ABAB) = P(AB)P(AB);$$

$$P(AbaB) = P(Ab)P(aB)$$

и т. п.

Гены, точнее их аллели, в гамете также встречаются случайно и независимо. Поэтому, если p, q — вероятности (частоты) генов A и a , а p_1, q_1 — вероятности (частоты) генов B и b , $p + q = 1$, $p_1 + q_1 = 1$, то $P(AB) = P(A) \cdot P(B) = pp_1$, $P(Ab) = P(A) \cdot P(b) = pq_1$, $P(aB) = P(a) \cdot P(B) = qp_1$, и $P(ab) = P(a) \cdot P(b) = qq_1$. Пользуясь этим, по исходным частотам генов p, q, p_1, q_1 можно рассчитать вероятности образования всех 16 генотипов с двумя признаками:

$$\begin{aligned} P(ABAB) &= P(AB) \cdot P(AB) = \\ &= P(A) \cdot P(B) \cdot P(A) \cdot P(B) = p^2 p_1^2; \end{aligned}$$

$$\begin{aligned} P(AbaB) &= P(Ab) \cdot P(aB) = \\ &= P(A) \cdot P(b) \cdot P(a) \cdot P(B) = pq_1 qp_1 \end{aligned}$$

и т. п.

Если гетерозиготы не связаны с полом, то генотипы $A_1 = abAB$, $A_2 = aBAb$, $A_3 = AbaB$ и $A_4 = ABab$ неразличимы. (В табл. 4 эти генотипы стоят на «побочной диагонали» и записаны в форме, где в пары объединены аллели одного гена). Поскольку образования этих генотипов — попарно несовместные события, то по формуле сложения вероятностей получаем

$$\begin{aligned} P(A_1 + A_2 + A_3 + A_4) &= \\ &= P(A_1) + P(A_2) + P(A_3) + P(A_4) = 4qq_1pp_1. \end{aligned}$$

Эти же результаты можно получить, трактуя образование генотипа $ABAB$ как встречу генотипов AA и

BB , а генотипа $AbaB$ — как встречу генотипов Aa и bB . В самом деле,

$$P(ABAB) = P(AABB) = P(AA)P(BB) = p^2p_1^2,$$

и, если гетерозиготы не связаны с полом, то

$$\begin{aligned} P(AbaB) &= P(AabB) = P(Aa)P(bB) = \\ &= 2pq \cdot 2p_1q_1 = 4qq_1pp_1. \end{aligned}$$

Если какой-то из аллелей доминантен, то по формулам сложения и умножения можно найти и вероятность фенотипов. Например, если A и a — это гены кареглазости и голубоглазости, а B и b — гены курчавости и прямоволосости, то кареглазому и курчавому фенотипу соответствуют четыре несовпадающих генотипа: $AABB$, $AAbB$, $aABB$ и $aAbB$ (первый столбец табл. 4). Поэтому вероятность появления такого фенотипа равна

$$\begin{aligned} P &= P(AABB + AAbB + aABB + aAbB) = \\ &= P(AABB) + P(AAbB) + P(aABB) + P(aAbB) = \\ &= P(AA)P(BB) + P(AA)P(bB) + P(aA)P(BB) + \\ &\quad + P(aA)P(bB) = p^2p_1^2 + p^2 \cdot 2p_1q_1 + 2pqp_1^2 + \\ &\quad + 2pq \cdot 2p_1q_1 = p^2p_1(p_1 + 2q_1) + 2pqp_1(p_1 + 2q_1) = \\ &= p_1(1 + q_1)p(1 + q) = (1 - q^2)(1 - q_1^2). \end{aligned}$$

В частном случае все четыре исходных варианта гамет AB , Ab , aB и ab могут быть равновероятными. Тогда $pp_1 = pq_1 = qp_1 = qq_1$. Из этих равенств вместе с обычными равенствами $p + q = 1$, $p_1 + q_1 = 1$ получаем $p = q = p_1 = q_1 = 1/2$. Такая ситуация складывается, например, когда исследуемая популяция является потомством от скрещивания гетерозигот по обоим признакам $ABab \times ABab$. (Об этом шла речь при рассмотрении табл. 4.) Тогда

$$P(ABAB) = p^2p_1^2 = \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{1}{16},$$

$$P(AbaB) = 4qq_1pp_1 = \frac{4}{16} = \frac{1}{4}$$

и т. д.

Мы пришли к тем же вероятностям, что и при анализе табл. 4 с помощью классического определения. Од-

нако на этот раз результат получен на основании предположения о независимости передачи двух признаков потомству. Совпадение результатов говорит о справедливости такого предположения. Впервые независимость передачи признаков установил Г. Мендель, и этот принцип носит название второго закона Менделя.

Аналогично рассматриваются гаметы и генотипы с n признаками. При этом некоторые общие равенства теории сложения и умножения событий служат основой для введения новых биологических понятий. Так, например, если p_i — частоты (вероятности) генов A_i , то гетерозиготностью (точнее, мерой гетерозиготности) называется число $1 - \sum_{i=1}^n p_i^2$. Очевидно, что p_i^2 — это частота пары $A_i A_i$ (вероятность встречи A_i с A_i); сумма $\sum_{i=1}^n p_i^2$ — частота (вероятность) образования всех гомозигот, а $1 - \sum_{i=1}^n p_i^2$ — вероятность противоположного события — образования гетерозигот.

3. Вероятность цепи

Вероятность встречи, которую мы использовали в предыдущем разделе, — частный случай вероятности совмещения независимых событий. Часто ситуации совмещения событий удобно изображать схематически в виде цепи последовательно соединенных звеньев. в виде цепи последовательно соединенных звеньев.

Цепь приборов. Рассмотрим участок электрической цепи, содержащий два последовательно соединенных прибора: A и B (рис. 8, а).

Предположим, что приборы работают независимо один от другого, и каждый из них может либо пропустить ток (прибор исправен), либо не пропустить (прибор неисправен). Обозначим $P(A)$ и $P(B)$ вероятности исправности приборов A и B соответственно. Для того чтобы по участку цепи прошел ток, нужно,

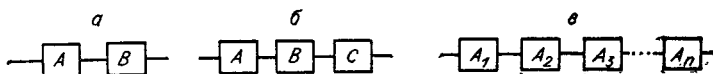


Рис. 8.

чтобы и прибор A , и прибор B были исправны, т. е. нужно совмещение исправности приборов. Так как приборы работают независимо, то по формуле умножения вероятностей вероятность прохождения тока выразится произведением

$$P = P(A) \cdot P(B). \quad (1)$$

Совершенно аналогично для трех последовательно соединенных и независимо работающих приборов A , B , C (рис. 8, б) вероятность прохождения тока по участку цепи выразится произведением

$$P = P(A) \cdot P(B) \cdot P(C),$$

а для n приборов A_1, A_2, \dots, A_n — произведением

$$P = P(A_1) \cdot P(A_2) \dots P(A_n).$$

В частности, если приборы однотипны, точнее говоря, если вероятности их исправности равны $P(A_1) = P(A_2) = \dots = P(A_n) = p$, то вероятность прохождения тока $P = p^n$.

Можно поставить в некотором смысле обратную задачу. Предположим, что вероятность исправности первого прибора $P(A)$ известна. После испытаний установили вероятность прохождения тока по всему участку P . Тогда из формулы (1) можно найти вероятность исправности второго прибора $P(B)$. Например, если $P(A) = 0,9$; $P = 0,72$, то в силу (1) $P(B) = P/P(A) = 0,72/0,9 = 0,8$.

В рассмотренном примере цепь и ее звенья представляют собой материальные объекты — провода, приборы и т. п. В общем случае это не обязательно так. Звеньями могут быть математические формулы, мнения людей, химические реакции, биологические признаки и т. п. Рассмотрим несколько примеров подобного рода.

Цепь инстанций. Предположим, что несколько инстанций или руководителей A_1, A_2, \dots, A_n последовательно подчинены один другому (рис. 8, в). Предварительное знакомство каждого из них с некоторым проектом показало, что каждый из руководителей на 90 % согласен утвердить проект. Какова вероятность p , что проект будет утвержден при последовательном прохождении по цепи инстанций?

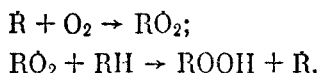
Обозначив $P(A_i)$ — вероятность утверждения проекта i -й инстанцией, по формуле умножения вероятностей можем написать

$$p = P(A_1) P(A_2) \dots P(A_n) = \underbrace{0,9 \cdot 0,9 \dots 0,9}_{n \text{ раз}} = 0,9^n.$$

Если $n = 5$, то $p = 0,9^5 \approx 0,53$, т. е. вероятность утверждения проекта немного больше половины. Если же $n = 10$, то $p = 0,9^{10} \approx 0,35$, т. е. вероятность утверждения снизилась почти до одной трети.

Это, конечно, не совсем корректный пример, так как речь идет об одном проекте, а не о серии однотипных проектов. Тем не менее он наглядно показывает, что чем длиннее цепь, тем менее вероятна ее безотказная работа, даже если каждое ее звено работает хорошо.

Цепь реакций. Цепной называют химическую реакцию, которая представляет собой цепочку одинаковых звеньев. Звеном может быть одна, две, реже — несколько стадий. Например, звено



начавшись с появлением свободного радикала углеводорода \dot{R} , во второй стадии снова выделяет этот радикал и тем самым создает возможность повторения такого же звена.

На некотором этапе цепная реакция может обрываться. Причиной обрыва может служить захват свободного радикала стенкой сосуда, действие ингибитора и т. п. Таким образом, на каждом этапе существует некоторая вероятность p продолжения цепи и вероятность $q = 1 - p$ обрыва цепи.

Какова вероятность, что цепная реакция содержит n звеньев? Для осуществления такой реакции нужно, чтобы n раз произошло продолжение реакции и после этого произошел обрыв. Так как процессы продолжения и обрыва независимы, то по формуле умножения вероятностей для $P(n)$ — вероятности появления цепи длины n — можем написать

$$P(n) = \underbrace{p \cdot p \dots p}_{n \text{ раз}} \cdot q = p^n q = p^n (1 - p).$$

Молекула полимера. Процесс полимеризации состоит в том, что к звену-мономеру присоединяется такой же мономер, к этому звену — еще один такой же мономер и т. д. Присоединение происходит с некоторой вероятностью p и, следовательно, не происходит с вероятностью $q = 1 - p$. Так как каждое следующее присоединение происходит независимо от предыдущих, то вероятность образования молекулы, содержащей n мономеров, как и в предыдущем примере, вычисляется по формуле

$$P(n) = \underbrace{p \cdot p \dots p}_{n \text{ раз}} \cdot q = p^n q = p^n (1 - p).$$

Вероятность возраста. Пусть l_n — число жителей данной страны или региона, доживших до возраста n лет. Тогда отношение $p_n = l_{n+1}/l_n$ называют вероятностью дожития от n до $n+1$ лет. Эту вероятность определяют статистически. Она зависит от многих причин и время от времени пересчитывается в каждой стране заново.

Предположим, что нам известны $p_1, p_2, \dots, p_n, \dots, p_\omega$, где ω — предельный возраст. Найдём вероятность $p_n(m)$ дожить от возраста n до возраста $n+m$ лет и умереть в следующем после этого году. Очевидно, чтобы такое событие произошло, нужно дожить от n до $n+1$ лет, затем — от $n+1$ до $n+2$ лет и т. д., наконец, от $n+m-1$ до $n+m$ лет и умереть в следующем после этого году. Таким образом, дожитие точно до $n+m$ лет — это произведение событий, и по формуле умножения вероятностей можем написать

$$p_n(m) = p_n p_{n+1} \dots p_{n+m-1} q_{n+m},$$

где

$$q_{n+m} = 1 - p_{n+m}.$$

Эта формула вполне аналогична формуле для вероятности длины цепной реакции или длины молекулы. Разница в том, что вероятность продолжения цепи дожития своя для каждого звена.

В частности, при $n=1$ получаем формулу вероятности дожития от годовалого возраста до возраста m :

$$p_1(m) = p_1 p_2 \dots p_m q_{1-m}.$$

Подчеркнем, что мы вычисляли вероятность дожития точно до возраста $n+m$, т. е. вероятность продол

жения цепи до возраста $n + m$ и обрыва цепи на этом звене. Если требование обрыва снять, т. е. вычислять вероятность дожития по крайней мере до возраста $n + m$, то мы бы получили

$$\tilde{p}_1(m) = p_1 p_2 \dots p_m.$$

Обратим внимание на то, что из этого равенства с очевидностью следует парадоксальный на первый взгляд вывод: вероятность дожития годовалого ребенка до 100 лет $\tilde{p}_1(100) = p_1 p_2 \dots p_{99} p_{100}$ меньше, чем вероятность p_{100} , т. е. вероятность дожития 99-летнего старца до 100 лет.

4. Дублирование

Цепь последовательно соединенных приборов, событий, объектов — один из крайних, наиболее простых типов соединений. Другим простейшим типом является параллельное соединение.

Начнем опять с приборов. Рассмотрим участок цепи, содержащий два прибора A и B , соединенных параллельно (рис. 9). Предположим, что приборы работают независимо и $P(A)$ — вероятность прохождения сигнала по прибору A , а $P(B)$ — по прибору B . Например, сигнал проходит по прибору, если прибор исправен, и не проходит — в противном случае. Очевидно, сигнал пройдет, если будет исправен хотя бы один прибор. Таким образом, вероятность прохождения сигнала по участку цепи — это вероятность $P(A + B)$, где сумма $A + B$ означает исправную работу хотя бы одного из приборов. Так как приборы работают независимо, то эту вероятность можно вычислить по

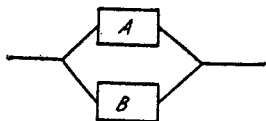


Рис. 9.

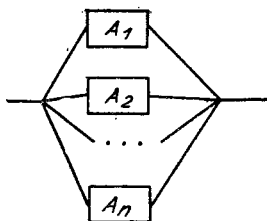


Рис. 10.

формуле

$$P(A + B) = P(A) + P(B) - P(A)P(B). \quad (2)$$

Например, если $P(A) = 0,8$, $P(B) = 0,9$, то

$$P(A + B) = 0,8 + 0,9 - 0,8 \cdot 0,9 = 0,98. \quad (3)$$

Можно поставить и обратную задачу. Предположим, что один из приборов — эталонный и вероятность его безотказной работы (т. е. вероятность прохождения по нему сигнала) известна. После испытаний установили вероятность прохождения сигнала по всему участку. Тогда из формулы (2) можно найти вероятность безотказной работы второго прибора. Например, если $P(A) = 0,8$, $P(A + B) = 0,95$, то, подставив это в (2), будем иметь

$$0,95 = 0,8 + P(B) - 0,8 \cdot P(B).$$

Отсюда

$$P(B) = \frac{0,15}{0,2} = 0,75.$$

Из формулы (3) видно, что параллельное соединение увеличило вероятность прохождения сигнала. Вторым прибор подстраховывает, дублирует первый. Можно ожидать, что параллельное соединение трех и более приборов еще более увеличит эту вероятность.

Если участок цепи состоит из n независимо работающих приборов, соединенных параллельно (рис. 10), и A_i означает, что сигнал прошел по i -му прибору, т. е. что i -й прибор исправен, то вероятность прохождения сигнала по участку — это вероятность исправной работы хотя бы одного прибора, т. е. вероятность суммы $A_1 + A_2 + \dots + A_n$. Такую вероятность проще всего подсчитать по четвертой формуле сложения:

$$P(A_1 + A_2 + \dots + A_n) = 1 - P(\bar{A}_1) \cdot P(\bar{A}_2) \dots P(\bar{A}_n). \quad (4)$$

Так как $P(\bar{A}_i) = 1 - P(A_i) < 1$, то при большом n произведение $P(\bar{A}_1)P(\bar{A}_2)\dots P(\bar{A}_n)$ достаточно мало и, следовательно, вероятность $P(A_1 + A_2 + \dots + A_n)$ близка к единице.

В частности, если все вероятности равны $P(A_1) = P(A_2) = \dots = P(A_n) = p$, то $P(\bar{A}_i) = 1 - P(A_i) = 1 - p = q$ и из формулы (4) получаем

$$P(A_1 + A_2 + \dots + A_n) = 1 - q^n. \quad (5)$$

Видно, что даже при малой вероятности p , т. е. при q , близкой к единице, выбирая достаточно большое n , можно сделать вероятность $P(A_1 + \dots + A_n)$ достаточно близкой к единице.

В параллельном соединении могут работать не только приборы. Вернемся, например, к задаче о стрельбе по мишени из двух орудий. В первой главе мы нашли вероятность поражения мишени, пользуясь классической схемой, шарами и урнами. Намного проще это сделать, если воспользоваться формулами (2) и (3).

Вместо схемы с параллельно соединенными приборами или одновременно стреляющими орудиями можно рассмотреть участок нервной или кровеносной системы, систему химических реакций, транспортную или промышленную схему с поставщиками, потребителями и т. п. Из формул (4) и (5), как мы отмечали, следует возможность получить большую вероятность суммы событий даже в том случае, когда каждое отдельное событие маловероятно. Для этого нужно, чтобы событий было достаточно много. Образно говоря, если к цели можно прийти многими параллельными путями, то вероятность достижения цели оказывается достаточно большой даже в том случае, когда каждый путь маловероятен. Для этого надо, чтобы путей было достаточно много. Можно сказать также, что система из большого числа параллельно соединенных звеньев может работать достаточно надежно, даже если каждое звено не обладает высокой надежностью. Это и есть принцип дублирования. Проявление этого принципа легко обнаружить в химических превращениях, в работе экспериментальных и промышленных установок. Он играет одну из главных ролей в живой природе. Вспомним, например, о полинуклеотидных молекулах с многократно повторяющимися одинаковыми локусами, о кодировании одной и той же аминокислоты различными триплетами — кодонами, о диглобдине и полиглобдине, о параллельных колбочках в глазу, о неслетном количестве икринок, нервных окончаний или клеток мозга, дублирующих друг друга.

5. Последовательные и параллельные соединения

В предыдущих пунктах мы рассмотрели порознь последовательные и параллельные соединения приборов и установили, как вычисляется вероятность прохождения сигнала по участку схемы в том и другом случае. На практике приходится иметь дело с различными сочетаниями соединений обоих типов. Рассмотрим два характерных примера.

Предположим, что сигнал проходит по участку схемы, состоящему из двух параллельных блоков A и B , первый из которых состоит из одного прибора A , а второй содержит два последовательно соединенных прибора B_1 и B_2 (рис. 11, a). Пусть возможность отказа одного из приборов не зависит от работы остальных. Сигнал проходит, если хотя бы один из блоков исправен, а каждый из блоков выходит из строя, если хотя бы один из его приборов отказал.

Обозначим $P(A)$, $P(B_1)$ и $P(B_2)$ — вероятности безотказной работы соответствующих приборов; $P(B)$ — вероятность исправности блока B (вероятность исправности блока A , очевидно, равна $P(A)$); $P(A+B)$ — вероятность прохождения сигнала по цепи. Тогда, используя формулы сложения и умножения, можем написать

$$\begin{aligned} P(A+B) &= P(A) + P(B) - P(A) \cdot P(B) = \\ &= P(A) + P(B_1) \cdot P(B_2) - P(A) \cdot P(B_1) \cdot P(B_2). \end{aligned} \quad (6)$$

На практике чаще задается не вероятность безотказной работы, а вероятность отказа, т. е. $P(\bar{A})$, $P(\bar{B}_1)$, $P(\bar{B}_2)$ и т. п. Так как отказ и безотказная работа — взаимно противоположные события, то $P(A) = 1 - P(\bar{A})$; $P(B_1) = 1 - P(\bar{B}_1)$; $P(B_2) = 1 - P(\bar{B}_2)$ и мы снова можем применить формулу (5).

Например, если $P(\bar{A}) = 0,1$; $P(\bar{B}_1) = 0,07$; $P(\bar{B}_2) = 0,08$, то $P(A) = 0,9$; $P(B_1) = 0,93$; $P(B_2) = 0,92$. По-

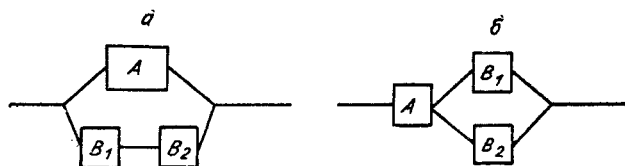


Рис. 11.

этому $P(A + B) = 0,9 + 0,93 \cdot 0,92 - 0,9 \cdot 0,93 \cdot 0,92 = 0,986$.

Теперь предположим, что участок схемы состоит из двух последовательно соединенных блоков A и B , один из которых состоит из одного прибора A , а другой содержит два параллельно соединенных прибора B_1 и B_2 (рис. 11, б). Пусть по-прежнему приборы работают независимо. Блок B выходит из строя, если отказали оба его прибора. Сигнал проходит, если оба блока A и B исправны. Обозначив $P(AB)$ — вероятность прохождения сигнала по цепи и сохранив остальные обозначения для вероятностей, можем написать $P(AB) = P(A) \cdot P(B) = P(A) [P(B_1) + P(B_2) - P(B_1) \cdot P(B_2)]$. В частности, для данных предыдущего примера

$$P(AB) = 0,9(0,93 + 0,92 - 0,93 \cdot 0,92) = 0,895.$$

Аналогично рассматриваются случаи, когда блок A состоит из двух или большего числа приборов, соединенных параллельно или последовательно, когда блоков более чем два и т. п.

Разумеется, и в этом случае вместо приборов могут быть рассмотрены химические реакции или установки, стреляющие устройства, нервные волокна, особи или популяции, предприятия и поставщики и т. п.

6. Задача о числе дублеров

Эта задача в зависимости от времени и конкретного содержания называлась по-разному. Но математическая ее основа — одна. В известном смысле эта задача является обратной по отношению к задаче о дублировании.

Вернемся к формуле (5). Если событие A состоит в том, что наступает по крайней мере одно из n независимых событий A_i и $P(A_i) = p$ для всех i , то вероятность такого события A , т. е. вероятность суммы $A_1 + A_2 + \dots + A_n$, мы вычисляли по формуле

$$P(A) = P(A_1 + A_2 + \dots + A_n) = 1 - q^n, \quad q = 1 - p. \quad (7)$$

Таким образом, по заданным значениям p и q вычисляется неизвестная вероятность $P(A)$.

В обратной задаче вероятность $P(A)$ известна, и нужно установить, при каком числе n независимых событий A_i достигается указанное значение $P(A)$. Точнее говоря, задается некоторый порог Q , ниже которого вероятность $P(A)$ не может опуститься, а вычисляется n — число дублеров, при котором

$$P(A) = 1 - q^n \geq Q. \quad (8)$$

Задача де Мере. Самой первой задачей о числе дублеров была одна из задач, предложенных кавалером де Мере Паскалю: «Сколько раз нужно подбросить две игральные кости, чтобы вероятность выпадения хотя бы один раз двух шестерок была бы больше $1/2$ ».

Событием A_i является выпадение двух шестерок при i -м подбрасывании. Значение порога $Q = 1/2$. Найдем вероятность $p = P(A_i)$.

Так как с каждой из шести граней первой кости может выпасть любая из шести граней второй кости, то всего равновозможных и попарно несовместных событий $6 \cdot 6 = 36$. Только одно из них — выпадение шестерки и на первой, и на второй кости — благоприятствует событию A_i . Таким образом, $P(A_i) = p = 1/36$. Отсюда $q = 1 - p = 1 - 1/36 = 35/36$.

Подбрасывания костей — очевидно, независимые испытания. Поэтому можем воспользоваться формулой (7), а следовательно, и (8). В нашем случае имеем

$$1 - \left(\frac{35}{36}\right)^n > \frac{1}{2}, \text{ т. е. } \left(\frac{35}{36}\right)^n < \frac{1}{2}.$$

Из этого неравенства и нужно найти n .

Эта задача решается совсем просто. Логарифмируя, получим $n \ln \frac{35}{36} < \ln \frac{1}{2}$. Отсюда

$$n > \frac{\ln 2}{\ln 36 - \ln 35} = \frac{0,6931}{0,0284} = 24,4.$$

Таким образом, чтобы вероятность выпадения двух шестерок была больше $1/2$, нужно подбросить кости не менее 25 раз.

Задача о стрельбе по самолету. В двадцатом веке с его окопными войнами возник другой вариант задачи о числе дублеров. Было замечено, что

одиночные выстрелы по низко летящему самолету успеха не приносят. Но если одновременно стреляют много стрелков, то довольно часто самолет сбивается. Каким должно быть число стрелков n , чтобы вероятность поражения самолета была не меньше 0,9?

В этом случае A_i — поражение самолета отдельным стрелком, $A = A_1 + A_2 + \dots + A_n$ — поражение самолета хотя бы одним из n стрелков. Обозначив $p = P(A_i)$, $q = 1 - p$ и учтя, что стрельба всеми стреляющими ведется независимо, можем воспользоваться формулами (7) и (8). Значение порога $Q = 0,9$. Таким образом, n удовлетворяет неравенству

$$1 - q^n \geq 0,9. \quad (9)$$

Отсюда $n \lg q \leq \lg 0,1$ и, следовательно,

$$n \geq -\frac{1}{\lg q}$$

($q < 1$, поэтому $\lg q < 0$). Из этого неравенства заключаем, что если, например, $p = 0,01$, то $q = 0,99$ и, следовательно,

$$n \geq -\frac{1}{\lg 0,99} = \frac{10000}{44} \approx 227.$$

Таким образом, если стреляет один стрелок, самолет поражается в одном случае из ста ($p = 0,01$). Если же одновременно стреляют 228 стрелков, то можно ожидать поражения самолета примерно в 9 случаях такой стрельбы из 10.

Такие же рассуждения и расчеты используются в радиационной биологии и химии. Мишенями служат участки молекул, хромосом или клеток, а снарядами — радиоактивные частицы.

Число проб. Задачи о числе дублеров особенно интересны, когда вероятность успеха p в отдельном испытании мала, как в случае стрельбы по самолету, но за счет увеличения числа дублирующих друг друга испытаний вероятность успеха в их сумме $P(A)$ удастся сделать значительной. Эта идея используется в физике при открытии новых элементарных частиц. Подобных задач немало и в биологии.

Например, раковина с жемчужиной встречается с частотой 1 на 1000; плодовая мушка с красными глазами — с частотой 3 на 100; мутантный ген фиксиру-

ется в одном случае из миллиона, каждый сотый человек — дальтоник и каждый двадцатый обладает редкой группой крови АВ. Сколько нужно поднять раковин, какой должен быть объем популяции, сколько мушек и людей нужно взять для анализа, чтобы с вероятностью 0,9 среди отобранных оказался хотя бы один экземпляр с указанным свойством? Мы сформулировали пять задач в одной фразе, так как все они решаются единой формулой (9). В качестве p нужно взять соответственно 0,001; 0,03; 10^{-6} ; 0,01 и 0,05. Значение порога Q мы положим равным 0,9. Разумеется, можно было бы взять другие значения и использовать формулу (8).

Как планировать семью. Предположим для простоты, что вероятности рождения мальчика и девочки равны. Обозначим их p и q . Таким образом, $p = q = 1/2$. Сколько нужно планировать детей, чтобы вероятность иметь хотя бы одного мальчика была более 0,9?

И эта задача решается по схеме задачи о дубле-рах. Событие A_i — рождение мальчика, событие \bar{A}_i — рождение девочки, $P(A_i) = p = 1/2$, $P(\bar{A}_i) = q = 1/2$. Рождение хотя бы одного мальчика после n рождений — это событие $A = A_1 + A_2 + \dots + A_n$. Поскольку события A_i мы считали независимыми, можем снова воспользоваться формулами (7) и (8), положив $Q = 0,9$.

Имеем

$$1 - \left(\frac{1}{2}\right)^n > 0,9.$$

Отсюда $(1/2)^n < 0,1$. Легко видеть, что это неравенство не выполняется при $n = 3$, но становится справедливым уже при $n = 4$. Таким образом, с вероятностью 0,9 в семье с четырьмя детьми имеется хотя бы один мальчик.

Антитела атакуют вирус. Вернемся к задаче, о которой мы говорили в разделе «Геометрическая вероятность». Поверхность вируса — сферу радиуса R — атакуют антитела — тонкие цилиндрики, которые, прикрепляясь к вирусу торцами, образуют вокруг точки прикрепления некоторую зону — тень площади S . Зоны могут перекрываться. Точками, находящимися

в тени, вирус не может прикрепиться к здоровой клетке. Таким образом, каждая атака A_i с вероятностью $p = S/4\pi R^2$ лишает вирус возможности прикрепиться к клетке. (Напомним, что $4\pi R^2$ — это площадь сферы радиуса R .) Естественно поставить вопрос: сколько антител должно одновременно участвовать в атаке, чтобы, например, на 90 % лишить вирус возможности прикрепления?

Атака n антител — это событие $A = A_1 + A_2 + \dots + A_n$. Результатом события A является лишение вируса возможности прикрепиться точками тени, образованной n антителами. Вероятность образования такой тени равна $P(A) = P(A_1 + A_2 + \dots + A_n)$. Так как антитела атакуют независимо, то $P(A) = P(A_1 + A_2 + \dots + A_n) = 1 - P(\bar{A}_1)P(\bar{A}_2) \dots P(\bar{A}_n) = 1 - q^n$, где $q = 1 - p$. По условию $P(A) = 0,9$. Таким образом, $1 - q^n = 0,9$. Отсюда, как и ранее, находится n .

Еще один вариант задачи о дублерах. Этот вариант состоит в том, что вместе с порогом Q задается не p (или q), а n . Так бывает, когда число дублеров по тем или иным причинам ограничено. Нужно найти такое p , чтобы вероятность $P(A)$ по-прежнему была не меньше некоторого порога Q .

Например, парашютист может нести на себе только два парашюта. Какой должна быть вероятность раскрытия одного парашюта, чтобы вероятность раскрытия хотя бы одного парашюта была не меньше 0,999999?

Событием A_i является раскрытие i -го парашюта, $i = 1, 2$; $n = 2$. Требуется, чтобы

$$P(A_1 + A_2) \geq Q = 1 - 10^{-6}.$$

Обозначим $P(A_1) = P(A_2) = p$, $P(\bar{A}_1) = P(\bar{A}_2) = q = 1 - p$. Так как исправность одного из парашютов не зависит от состояния другого, то

$$P(A_1 + A_2) = 1 - P(\bar{A}_1)P(\bar{A}_2) = 1 - q^2.$$

По условию $1 - q^2 \geq 1 - 10^{-6}$. Отсюда $q^2 \leq 10^{-6}$, $q \leq 10^{-3}$ и, следовательно, $p \geq 1 - 10^{-3} = 0,999$.

Подобные задачи возникают при оснащении самолета или ракеты запасными двигателями, при оборудовании опасного участка блокираторами и вообще всякий раз, когда нужно с достаточно большой веро-

ятностью гарантировать наступление некоторого события при ограниченном числе дублеров.

Сборка полинуклеотидной молекулы. Закончим раздел задач из молекулярной биологии. Известно, что молекулы ДНК и РНК собираются по принципу комплементарности на матрицах — «негативных» молекулах: буква *A* узнает букву *T*, буква *G* — букву *C* и наоборот. Это узнавание происходит с разными вероятностями, но мы для простоты предположим, что вероятность узнавания всех букв одна и та же и равна q , а вероятность ошибки, следовательно, равна $p = 1 - q$.

Пусть A_i — ошибка в i -м месте. Если произойдет хоть одно событие A_i , молекула правильно не воспроизведется. Таким образом, вероятность невоспроизведения является вероятностью суммы $P(A) = P(A_1 + A_2 + \dots + A_n)$. Эту вероятность называют качеством передачи информации. Как и ранее, можем написать

$$P(A) = P(A_1 + A_2 + \dots + A_n) = \\ = 1 - P(\bar{A}_1)P(\bar{A}_2) \dots P(\bar{A}_n) = 1 - q^n.$$

Вероятность противоположного события \bar{A} является вероятностью правильного воспроизведения. Очевидно, что

$$P(\bar{A}) = 1 - P(A) = 1 - (1 - q^n) = q^n.$$

Этот результат ясен и из других соображений: чтобы произошла правильная сборка всего слова, нужно, чтобы на каждом i -м месте произошло узнавание, т. е. чтобы одновременно и независимо произошли события $\bar{A}_1, \bar{A}_2, \dots, \bar{A}_n$. Вероятность такого пересечения событий равна произведению их вероятностей:

$$P(\bar{A}) = P(\bar{A}_1 \bar{A}_2 \dots \bar{A}_n) = \\ = P(\bar{A}_1)P(\bar{A}_2) \dots P(\bar{A}_n) = q^n.$$

Так как воспроизведение молекулы — случайное событие, то существует некоторая вероятность, что на каком-то этапе процесс воспроизведения молекул оборвется. Произойдет слишком много ошибок на предыдущих этапах, и вероятность правильной сборки заметного числа молекул на следующих этапах окажется ничтожно малой. Произойдет, как говорят, катастро-

фа ошибок. Физико-химический анализ и анализ соответствующих дифференциальных уравнений процесса воспроизведения показывают, что вероятность продолжения процесса будет достаточно большой, если вероятность $P(\bar{A}) = q^n$ будет больше некоторого порога Q .

Таким образом, для непрерывного воспроизведения молекул необходимо, чтобы

$$q^n > Q.$$

Отсюда

$$n < \frac{\ln Q}{\ln q}.$$

Значения Q и q находят экспериментально, а из написанного неравенства определяют максимальную длину молекул, способных к правильному воспроизведению. Видно, что эта длина растет с ростом q (чем ближе q к 1, тем ближе $\ln q$ к нулю). В природе $q \approx 0,99$. В эксперименте с помощью ферментов удается довести q до значения 0,99999. Подробности можно найти в книге М. Эйгена «Самоорганизация материи и эволюция биологических макромолекул» (М.: Мир, 1973).

7. Условная и полная вероятность

Теоремы об условной и полной вероятности вместе с формулами Байеса дают возможность простыми средствами решать вполне содержательные задачи. Несколько задач такого рода мы рассмотрим в этом разделе. Начнем, как обычно, с самых простых задач.

Извлечение без возвращения. Одну задачу этого типа мы рассмотрели, когда вводили понятие условной вероятности. Задача, напомним, состояла в вычислении вероятности вынуть подряд два шара одного цвета при условии, что выпутый шар не возвращается. Подобные задачи легко обобщаются на случай более чем двух извлечений.

Рассмотрим, например, уже знакомую задачу о студенте, который знает ответ на 20 вопросов из 26. Предположим, что вопросы задаются последовательно один за другим. Найдем вероятность того, что три под-

ряд заданных вопроса — счастливые. Пусть A, B, C — события, состоящие соответственно в том, что первый, второй и третий вопросы — счастливые. Найдем $P(ABC)$. По общей формуле условной вероятности можем написать

$$\begin{aligned} P(ABC) &= P(AB) \cdot P(C/AB) \\ P(AB) &= P(A) \cdot P(B/A). \end{aligned} \quad (10)$$

Таким образом,

$$P(ABC) = P(A) \cdot P(B/A) \cdot P(C/AB). \quad (11)$$

Пользуясь классической схемой, находим $P(A) = 20/25$. После того как первый вопрос задан и оказалось, что он счастливый, осталось еще 24 вопроса, из них 19 счастливых. Поэтому по классической схеме $P(B/A) = 19/24$. Если же и второй вопрос оказался счастливым, то осталось 23 вопроса и 18 из них счастливых. Поэтому по классической схеме $P(C/AB) = 18/23$. Подставив найденные вероятности в (10), получим

$$P(ABC) = \frac{20}{25} \cdot \frac{19}{24} \cdot \frac{18}{23} = \frac{57}{115}.$$

Конструирование цепи из счастливых вопросов ABC напоминает конструирование цепи молекулы из мономеров. Мономер, вошедший в молекулу, также не возвращается в общий пул мономеров. Однако если мономеров очень много или если реакция открытая, мономеры поступают в реактор, то присоединение очередного мономера к молекуле полимера не меняет вероятности выбора мономера на следующем этапе. Поэтому вместо условной вероятности при рассмотрении конструкции полимера пользуются безусловной. Если же объектов немного и они не поступают извне, то каждое извлечение объекта, как мы видели, меняет вероятность удачного извлечения на следующем этапе. События становятся зависимыми, и нужно пользоваться условной вероятностью.

В примере со счастливыми вопросами объекты, которые нужно выбирать, обладают одним и тем же свойством. В общем случае это может быть не так. При сборке агрегата его детали, как правило, разнообразны, и их нужно выбирать в определенном поряд-

ке. Такая же ситуация при «сборке» молекул ДНК из нуклеотидов, молекул белка из аминокислот и т. п. Однако и в этом случае вероятность совмещения вычисляется по формулам (10) и (11). Рассмотрим простые примеры такого рода.

Предположим, что в ящике лежит 10 деталей, из которых 4 первого типа и 6 — второго. Для сборки агрегата нужно сначала взять деталь первого типа, а затем — второго. Какова вероятность, что при выборе наугад детали будут взяты в нужной последовательности?

Пусть A и B означают соответственно выбор деталей первого и второго типов. Тогда интересующее нас событие есть совмещение AB , и по формуле условной вероятности

$$P(AB) = P(A) \cdot P(B/A).$$

Пользуясь классической схемой, находим $P(A) = 4/10$. После того как деталь первого типа взята, в ящике осталось 9 деталей, из них 6 — второго типа. Поэтому по классической схеме $P(B/A) = 6/9$. Отсюда

$$P(AB) = \frac{4}{10} \cdot \frac{6}{9} = \frac{4}{15}.$$

Теперь предположим, что для сборки нужно взять сначала 2 детали первого типа, а затем одну второго. Обозначив A , B , C последовательный выбор деталей первого типа, первого и второго типа, можем применить формулу (11). При этом, пользуясь классической схемой, находим $P(A) = 4/10$, $P(B/A) = 3/9$, $P(C/AB) = 6/8$. Поэтому

$$P(ABC) = \frac{4}{10} \cdot \frac{3}{9} \cdot \frac{6}{8} = \frac{1}{10}.$$

Совершенно аналогично вычисляется вероятность сборки из деталей трех и большего числа типов. Пусть, например, в ящике лежит 10 деталей, из которых 5 первого типа, 3 — второго и 2 — третьего. Какова вероятность, что при выборе наугад первой будет взята деталь первого типа, второй — второго и третьей — третьего? Обозначив A , B , C последовательный выбор деталей соответствующего типа и пользуясь классической схемой, получим $P(A) = 5/10$, $P(B/A) =$

$= 3/9$, $P(C/AB) = 2/8$. Подставив это в формулу (11), найдем

$$P(ABC) = \frac{5}{10} \cdot \frac{3}{9} \cdot \frac{2}{8} = \frac{1}{24}.$$

Заменяв детали разных типов различными нуклеотидами или аминокислотами, получим задачи о вероятности сборки молекул ДНК или молекул белков.

Курящие мужчины и вакцинированные больные. В предыдущем пункте мы вычисляли вероятности произведений событий по известной или заранее вычисленной условной вероятности. Часто требуется решить обратную задачу — найти условную вероятность по заданной вероятности произведения. Такая ситуация возникает при выборе объекта, обладающего некоторым свойством, среди объектов, обладающих еще и другим каким-нибудь свойством, например, при выборе курящих среди мужчин, девочек среди детей, вакцинированных среди больных, мушек с мутацией глаз среди мушек, имеющих еще и мутацию крыльев, и т. п.

Если заданы $P(AB)$ и $P(A) \neq 0$, то, как известно,

$$P(B/A) = \frac{P(AB)}{P(A)}. \quad (12)$$

Этой формулой и пользуются при вычислении условной вероятности $P(B/A)$.

Например, в терапевтическом отделении больницы 70 % пациентов — женщины, а 21 % — курящие мужчины. Наугад выбирают пациента. Он оказывается мужчиной. Какова вероятность, что он курит?

Пусть M означает, что пациент — мужчина, а K — что пациент курит. Тогда по условию $P(M) = 0,3$, а $P(MK) = 0,21$. Поэтому по формуле (12) искомая условная вероятность

$$P(K/M) = \frac{P(MK)}{P(M)} = \frac{0,21}{0,3} = 0,7.$$

Аналогичная задача: в группе туристов 20 % детей, причем 12 % — девочки. Наугад выбирают ребенка. Какова вероятность, что это девочка? Какова вероятность, что это мальчик?

Пусть A означает, что турист — ребенок, $Ж$ — что турист женского пола, а $М$ — мужского. Тогда по ус-

ловию $P(A) = 0,2$, $P(KA) = 0,12$, $P(MA) = 0,08$. Следовательно,

$$P(K/A) = \frac{P(KA)}{P(A)} = \frac{0,12}{0,2} = 0,6,$$

$$P(M/A) = \frac{P(MA)}{P(A)} = \frac{0,08}{0,2} = 0,4.$$

Очевидно, что $P(K/A) + P(M/A) = P((K + M)/A) = P(U/A) = 1$, как это и положено для достоверного события.

Еще одна задача на медицинскую тему. В большом городе 90 % населения вакцинировано против гриппа. Тем не менее 20 % населения заболело, причем 4,5 % заболевших были до болезни вакцинированы.

Пусть B означает, что житель города вакцинирован, а E — что он болен. Каковы вероятности $P(E/B)$ и $P(B/E)$? Так как по условию $P(B) = 0,9$, $P(E) = 0,2$, $P(EB) = 0,045$, по формуле (12) имеем

$$P(E/B) = \frac{P(EB)}{P(B)} = \frac{0,045}{0,90} = 0,05.$$

Таким образом, частота больных среди вакцинированных в четыре раза меньше, чем $P(E) = 0,2$ — частота больных среди всего населения.

Аналогично,

$$P(B/E) = \frac{P(EB)}{P(E)} = \frac{0,045}{0,2} = 0,225.$$

Таким образом, частота вакцинированных среди больных в четыре раза меньше, чем $P(B) = 0,9$ — частота вакцинированных среди всего населения.

Равенство коэффициентов уменьшения, конечно, не случайное. Из известных формул $P(EB) = P(B) \times P(E/B)$, $P(EB) = P(E) \cdot P(B/E)$ получаем

$$\frac{P(E/B)}{P(B)} = \frac{P(B/E)}{P(E)} = \frac{P(EB)}{P(B) \cdot P(E)}.$$

В данном случае это отношение равно $1/4$.

В приведенных примерах множество рассматриваемых объектов разбивалось на два взаимно дополняющих подмножества: мужчины — женщины, больные — здоровые и т. п. В других ситуациях подмножеств может быть более двух. Например, в популяции плодовых мушек 30 % имеют мутацию глаз (Γ), 20 % —

мутацию крыльев (K) и 15 % — обе мутации ($ГK$). Какова вероятность, что у наугад выбранной мушки с мутацией глаз имеется и мутация крыльев? Какова вероятность, что у наугад выбранной мушки с мутацией крыльев имеется и мутация глаз?

По условию $P(Г)=0,3$, $P(K)=0,2$, $P(ГK)=0,15$. Поэтому

$$P(K/Г) = \frac{P(ГK)}{P(Г)} = \frac{0,15}{0,30} = 0,5.$$

$$P(Г/K) = \frac{P(ГK)}{P(K)} = \frac{0,15}{0,20} = 0,75.$$

Конкурирующие причины одного события. Часто встречаются ситуации, когда событие с различной вероятностью может иметь различные несовместные причины, а сами эти причины также носят случайный характер. В таких ситуациях полезной оказывается формула полной вероятности и связанная с ней формула Байеса. Рассмотрим несколько примеров.

Начнем со стандартной задачи о дефектных деталях. Предположим, что в приемном бункере вперемешку лежат 200 деталей от одного станка и 300 — от другого. Вероятность изготовления бракованной детали на первом станке равна 0,02, а на втором — 0,01. Какова вероятность, что наугад взятая из бункера деталь окажется бракованной?

Обозначим A_1 — событие, состоящее в том, что взятая наугад деталь изготовлена на первом станке, а A_2 — событие, состоящее в том, что деталь изготовлена на втором станке. Пусть B означает, что взятая деталь бракованная. Тогда по формуле полной вероятности можем написать

$$P(B) = P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2). \quad (13)$$

Вероятности $P(A_1)$ и $P(A_2)$ легко найти по классическому определению:

$$P(A_1) = \frac{200}{500} = \frac{2}{5}, \quad P(A_2) = \frac{300}{500} = \frac{3}{5}.$$

Условные вероятности заданы:

$$P(B/A_1) = 0,02, \quad P(B/A_2) = 0,01.$$

Подставив их в формулу (13), получим

$$P(B) = \frac{2}{5} \cdot 0,02 + \frac{3}{5} \cdot 0,01 = 0,014.$$

По схеме этой простой производственной задачи решаются многие задачи из биологии и медицины. Например, в санатории 30 % пациентов — мужчины (M) и 70 % — женщины ($Ж$). Сердечные больные среди мужчин встречаются в 2 раза чаще, чем среди женщин. Какова вероятность, что наугад выбранный пациент сердечник? Обозначив C — наличие сердечного заболевания, можем написать

$$P(M) = 0,3, P(Ж) = 0,7, P(C/M) = \frac{2}{3}, P(C/Ж) = \frac{1}{3}.$$

Подставив указанные значения в формулу полной вероятности, найдем $P(C)$.

Вместо пациентов разного пола можно рассмотреть пациентов, прибывших из разных областей, курящих и некурящих, соблюдающих режим и несоблюдающих, лечившихся ранее или нет и т. п.

Применим эту же схему к актуальной экологической задаче. На город примерно 100 дней в году дует ветер с севера и 200 дней в году — с запада. Промышленные предприятия, расположенные на севере, производят выброс вредных веществ каждый третий день, а расположенные на западе — в конце каждой недели. Как часто город подвергается воздействию вредных выбросов? Иными словами, какова вероятность, что в наугад выбранный день город будет накрыт промышленным смогом?

Обозначив C — ветер с севера, $З$ — ветер с запада и B — воздействие вредных выбросов на город, можем написать

$$P(C) = \frac{100}{365}, P(З) = \frac{200}{365}, P(B/C) = \frac{1}{3}, P(B/З) = \frac{1}{7}.$$

Отсюда по формуле полной вероятности

$$\begin{aligned} P(B) &= P(C) \cdot P(B/C) + P(З) \cdot P(B/З) = \\ &= \frac{100}{365} \cdot \frac{1}{3} + \frac{200}{365} \cdot \frac{1}{7} = 0,09 + 0,08 = 0,17. \end{aligned}$$

Таким образом, примерно два месяца в году город накрыт смогом.

А эта задача одинаково интересна и с медицинской, и с технической точки зрения. Известно, что некоторое заболевание имеется у 10 % населения. Контрольный тест распознает болезнь не всегда. Точнее говоря, он обнаруживает болезнь в 95 % случаев, когда она есть, и указывает иногда на ее наличие, когда ее нет. Непосредственное применение теста указало на наличие болезни у каждого 11 человек из 100. Какова вероятность, что больным будет признан здоровый человек?

Пусть B означает наличие болезни, \bar{B} — ее отсутствие, а Π — указание теста на наличие болезни. По условию $P(B) = 0,1$, $P(\bar{B}) = 0,9$. Нужно найти $P(\Pi/\bar{B})$. По формуле полной вероятности $P(\Pi) = P(B) \times P(\Pi/B) + P(\bar{B}) \cdot P(\Pi/\bar{B})$. Отсюда

$$P(\Pi/\bar{B}) = \frac{P(\Pi) - P(B) \cdot P(\Pi/B)}{P(\bar{B})}.$$

Подставив указанные вероятности, получим

$$P(\Pi/\bar{B}) = \frac{0,11 - 0,1 \cdot 0,95}{0,9} \approx 0,017.$$

Таким образом, 1,7 % здоровых людей при тестировании могут быть признаны больными.

Изменив постановку задачи, теми же средствами можно найти вероятность того, что данный, наугад взятый человек будет признан больным (или здоровым), что больной человек будет признан здоровым и т. п.

Рассмотрев вместо больных людей дефектные детали, мы узнаем, почему даже при 95 %-ной надежности приборов ОТК в продажу попадают изделия с браком.

До сих пор мы рассматривали только две конкурирующие причины случайного события. Это совсем не обязательное ограничение. Детали могут поступать не от двух, а от трех или четырех станков, пациенты в санаторий могут приехать из нескольких городов, а ветер на город может дуть по крайней мере с четырех сторон света.

Вероятности гипотез. Вернемся к задаче о станках и деталях. По формуле полной вероятности (13) мы нашли вероятность события B — вероятность дефекта в детали, которую собираемся взять наугад из бункера. Предположим теперь, что событие B на-

ступило. Деталь взята, обследована и в ней обнаружен дефект. Можно ли узнать, на каком станке изготовлена эта деталь. Очевидно, точного ответа дать нельзя. Можно только предположить с большей или меньшей вероятностью, что она изготовлена на первом или на втором станке, т. е. можно выдвинуть гипотезы о происхождении бракованной детали. Каковы вероятности этих гипотез, т. е. каковы вероятности $P(A_1/B)$ и $P(A_2/B)$? Ответить на этот вопрос нетрудно, если воспользоваться формулами Байеса:

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}, \quad (14)$$

где $P(B)$ вычисляется по формуле полной вероятности.

В нашем случае

$$P(A_1/B) = \frac{0,4 \cdot 0,02}{0,014} = \frac{4}{7}, \quad P(A_2/B) = \frac{0,6 \cdot 0,01}{0,014} = \frac{3}{7}.$$

Таким образом, до того как деталь извлечена из бункера, мы могли предсказать, что вероятность ее изготовления, например, на первом станке равна $P(A_1) = 2/5$. Но после того как деталь извлечена и на ней обнаружен брак (событие B), находим, что вероятность ее изготовления на первом станке равна $P(A_1/B) = 4/7$, что заметно отличается от $2/5$.

Применим формулы Байеса к решению некоторых медицинских и биологических задач.

Известно, что 25 % мужчин и 5 % женщин — дальтоники. В группе 18 мальчиков и 22 девочки. Какова вероятность, что наугад выбранный ребенок окажется мальчиком-дальтоником?

Обозначим M и D — события, состоящие в выборе соответственно мальчика и девочки, и C — наличие у выбранного ребенка дальтонизма. Тогда по условию $P(M) = 18/40$, $P(D) = 22/40$, $P(C/M) = 0,25$, $P(C/D) = 0,05$. Отсюда по формуле полной вероятности

$$\begin{aligned} P(C) &= P(M) \cdot P(C/M) + P(D) \cdot P(C/D) = \\ &= \frac{18}{40} \cdot 0,25 + \frac{22}{40} \cdot 0,05 = 0,14. \end{aligned}$$

и по формуле Байеса

$$P(M/C) = \frac{P(M) \cdot P(C/M)}{P(C)} = \frac{18}{40} \cdot \frac{1}{4} \cdot \frac{100}{14} = \frac{45}{56}.$$

Другая задача. Известно, что 1 из 700 мальчиков рождается с лишней Y -хромосомой и у этих мальчиков наблюдается агрессивное поведение в 20 раз чаще, чем у обычных. У данного мальчика агрессивное поведение. Какова вероятность, что у него лишняя Y -хромосома?

Пусть A означает наличие агрессивного поведения, Y — наличие лишней Y -хромосомы, а \bar{Y} — ее отсутствие. Тогда по условию $P(Y) = 1/700$, $P(\bar{Y}) = 699/700$, $P(A/Y) = 20/21$, $P(A/\bar{Y}) = 1/21$. Поэтому по формуле полной вероятности

$$\begin{aligned} P(A) &= P(Y) \cdot P(A/Y) + P(\bar{Y}) \cdot P(A/\bar{Y}) = \\ &= \frac{1}{700} \cdot \frac{20}{21} + \frac{699}{700} \cdot \frac{1}{21} = 0,05 \end{aligned}$$

и по формуле Байеса

$$P(Y/A) = \frac{P(Y) \cdot P(A/Y)}{P(A)} = \frac{1}{700} \cdot \frac{20}{21} \cdot \frac{100}{5} = 0,027.$$

Как видим, вероятность довольно мала. Поэтому далеко не каждого мальчика с агрессивным поведением надо зачислять в отряд обладателей лишней Y -хромосомы. Объясняется это тем, что хотя вклад обладателей лишней хромосомы в агрессивное поведение велик, но их доля среди поворожденных очень мала.

Рассмотрим еще актуальную задачу о допинге. Известно, что примерно 10 % классных спортсменов принимают перед соревнованиями допинг. Анализ тренировок данного спортсмена показывает, что без допинга он в шести попытках из девяти поднимает рекордный вес, а после допинга — в восьми попытках из десяти. На соревнованиях рекордный вес был взят. Какова вероятность, что спортсмен принял допинг?

Обозначим D — принятие допинга, \bar{D} — непринятие и B — взятие рекордного веса. Тогда по условию $P(D) = 0,1$, $P(\bar{D}) = 0,9$, $P(B/D) = 8/10$, $P(B/\bar{D}) = 6/9$. По формуле полной вероятности

$$\begin{aligned} P(B) &= P(D) \cdot P(B/D) + P(\bar{D}) \cdot P(B/\bar{D}) = \\ &= 0,1 \cdot \frac{8}{10} + 0,9 \cdot \frac{6}{9} = 0,68. \end{aligned}$$

Отсюда по формуле Байеса

$$P(D/B) = \frac{P(D) \cdot P(B/\bar{D})}{P(B)} = \frac{1}{10} \cdot \frac{8}{10} \cdot \frac{100}{68} = \frac{2}{17} \approx 0,12.$$

Этот результат заметно больше, чем общий показатель принятия допинга среди спортсменов (0,1). Это может служить основанием для привлечения спортсмена к анализу на допинг после соревнования.

Задачу можно разнообразить, рассмотрев не один, а несколько типов стимулирующих средств, или несколько режимов питания, или несколько комплексов тренировочных упражнений и т. п.

Рассмотрим, наконец, задачу о симптомах и болезнях. Предположим, что некоторый симптом S проявляется с различной частотой при заболеваниях M_1, M_2, \dots, M_n . Обозначим эти частоты $P(S/M_1), P(S/M_2), \dots, P(S/M_n)$. Регулярное санитарно-статистическое обследование позволяет установить частоту заболевания M_i в данной местности (например, число заболевших на 1000 жителей). Обозначим эти частоты $P(M_1), P(M_2), \dots, P(M_n)$. Приравня частоты за вероятности, можем по формуле полной вероятности найти $P(S)$ — вероятность проявления симптома S :

$$P(S) = P(M_1) \cdot P(S/M_1) + P(M_2) \cdot P(S/M_2) + \dots \\ \dots + P(M_n) \cdot P(S/M_n). \quad (15)$$

Теперь предположим, что у данного больного обнаружен симптом S и подозревается некоторое заболевание M_i . По формуле Байеса

$$P(M_i/S) = \frac{P(M_i) \cdot P(S/M_i)}{P(S)}.$$

Подставив в эту формулу известные значения $P(M_i), P(S/M_i)$ и вычисленное по формуле (15) значение $P(S)$, найдем вероятность заболевания M_i при условии обнаружения симптома S .

Например, у больного обнаружена повышенная температура (симптом S) и после первичного обследования, исключившего легко диагностируемые заболевания, подозреваются грипп (M_1), красная волчанка (M_2) и пневмония (M_3) (точнее говоря, некоторые варианты этих заболеваний). Известны вероятности

проявления симптома S при данных заболеваниях. Например,

$$P(S/M_1) = 0,4; \quad P(S/M_2) = 0,8; \quad P(S/M_3) = 0,4.$$

Распространенность заболеваний задается вероятностями

$$P(M_1) = 0,3; \quad P(M_2) = 0,05; \quad P(M_3) = 0,1.$$

Тогда по формуле полной вероятности

$$P(S) = 0,3 \cdot 0,4 + 0,05 \cdot 0,8 + 0,1 \cdot 0,4 = 0,2.$$

Отсюда по формуле Байеса

$$P(M_1/S) = \frac{0,3 \cdot 0,4}{0,2} = 0,6; \quad P(M_2/S) = \frac{0,05 \cdot 0,8}{0,2} = 0,2;$$

$$P(M_3/S) = \frac{0,1 \cdot 0,4}{0,2} = 0,2.$$

Таким образом, хотя при красной волчанке высокая температура проявляется значительно чаще, чем при данных формах гриппа и пневмонии, большая распространенность гриппа сделала его более вероятным. Полученные данные не могут, конечно, служить окончательным диагнозом. Но, опираясь на них, легче построить план дальнейшего обследования больного.

Рассмотренный пример показывает, как можно вычислять вероятности причин M_1, M_2, \dots, M_n (генотипов, сортов растений, пород животных, компонентов синтезируемых веществ и т. п.) по наблюдаемому признаку или свойству S .

Закон Харди — Вайнберга, хорошо известный генетикам, открыт в 1908 году. Но еще за пять лет до этого его установил Кастл. Речь идет о постоянном соотношении между частотами генотипов AA , Aa и aa , сохраняющемся из поколения в поколение в менделевской двуполюсной панмиктической популяции. Получим этот закон с помощью теорем о вероятностях.

Пусть частоты (вероятности) генов A и a в родительской популяции равны соответственно p и q . Тогда, как мы уже установили, вероятности генотипов AA , Aa и aa равны соответственно $P(AA) = p^2$, $P(Aa) = P(Aa + aA) = 2pq$ и $P(aa) = q^2$. Между мужскими и женскими генотипами возможны 9 типов

Таблица 5

Отец	Мать		
	AA	Aa	aa
AA	AA×AA	AA×Aa	AA×aa
Aa	Aa×AA	Aa×Aa	Aa×aa
aa	aa×AA	aa×Aa	aa×aa

Таблица 6

Тип скрещивания	Частота (вероятность скрещивания)	Потомство		
		AA	Aa	aa
$S_1 = AA \times AA$	$p^2 \cdot p^2 = p^4$	p^4	0	0
$S_2 = AA \times Aa$	$p^2 \cdot 2p \cdot q = 2p^3q$	p^3q	p^3q	0
$S_3 = AA \times aa$	$p^2 \cdot q^2 = p^2q^2$	0	p^2q^2	0
$S_4 = Aa \times AA$	$2pq \cdot p^2 = 2p^3q$	p^3q	p^3q	0
$S_5 = Aa \times Aa$	$2pq \cdot 2pq = 4p^2q^2$	p^2q^2	$2p^2q^2$	p^2q^2
$S_6 = Aa \times aa$	$2pq \cdot q^2 = 2pq^3$	0	pq^3	pq^3
$S_7 = aa \times AA$	$q^2 \cdot p^2 = p^2q^2$	0	p^2q^2	0
$S_8 = aa \times Aa$	$q^2 \cdot 2pq = 2pq^3$	0	pq^3	pq^3
$S_9 = aa \times aa$	$q^2 \cdot q^2 = q^4$	0	0	q^4
Частота генотипов в популяции потомства		p^2	$2pq$	q^2

скрещиваний, перечисленных в табл. 5. Вероятности этих типов скрещиваний, вычисленные как вероятности встреч, приведены во втором столбце табл. 6.

Найдем вероятности генотипов потомства. Для этого вычислим сначала вероятности этих генотипов при том или ином типе скрещивания. По формулам вероятности произведения можем написать

$$P(AA \cdot S_i) = P(S_i) \cdot P(AA/S_i);$$

$$P(Aa \cdot S_i) = P(S_i) \cdot P(Aa/S_i);$$

$$P(aa \cdot S_i) = P(S_i) \cdot P(aa/S_i).$$

Вероятности $P(S_i)$ указаны во втором столбце табл. 6, а условные вероятности мы в свое время находили по табл. 1. Пользуясь этим, будем иметь, например,

$$P(AA \cdot S_1) = P(S_1) \cdot P(AA/S_1) = p^4 \cdot 1 = p^4;$$

$$P(Aa \cdot S_2) = P(S_2) \cdot P(Aa/S_2) = 2p^3q \cdot \frac{1}{2} = p^3q.$$

Подобным образом находятся и другие вероятности. Их значения приведены в трех последних столбцах табл. 6.

Чтобы получить вероятность генотипа в потомстве, нужно сложить все вероятности, с которыми данный генотип получается при том или ином скрещивании, т. е. сложить вероятности внутри соответствующего столбца. При этом совершенно ясно, что мы пользуемся формулой полной вероятности. Например,

$$P(AA) = P(AA \cdot S_1) + P(AA \cdot S_2) + \dots + P(AA \cdot S_8) + \\ + P(AA \cdot S_9) = P(S_1) \cdot P(AA/S_1) + P(S_2) \cdot P(AA/S_2) + \\ + P(S_4) \cdot P(AA/S_4) + P(S_5) \cdot P(AA/S_5).$$

Остальные слагаемые мы не написали, так как условные вероятности образования генотипа AA при соответствующих типах скрещивания равны нулю.

Подставив в это равенство значения вероятностей, взятые из третьего столбца табл. 6, получим

$$P(AA) = p^4 + p^3q + p^3q + p^2q^2 = p^2(p^2 + 2pq + q^2) = \\ = p^2(p + q)^2 = p^2,$$

так как $p + q = 1$.

Аналогично, суммируя вероятности в столбцах для Aa и aa , получим

$$P(Aa) = 2(p^3q + 2p^2q^2 + pq^3) = 2pq(p^2 + 2pq + q^2) = \\ = 2pq;$$

$$P(aa) = p^2q^2 + 2pq^3 + q^4 = q^2(p^2 + 2pq + q^2) = q^2.$$

Таким образом, вероятности генотипов в популяции потомства точно такие же, как и в популяции родителей. В этом и состоит закон Харри — Вайнберга.

Если же различать так называемые реципрокные типы с одинаковыми генотипами, например $AA \times aa$ и $aa \times AA$, то табл. 6 заметно сокращается. Нужно объединить типы S_2 и S_4 , S_6 и S_8 , S_3 и S_7 , сложив соответствующие вероятности во втором столбце.

Кроме того, из табл. 6 можно получить важную информацию, связанную с полом родителей. Пусть, например, требуется найти частоты (вероятности) генотипов в потомстве в зависимости от генотипа отца. Генотип отца AA представлен в первых трех строчках

Отец	Ребенок			Сумма
	AA	Aa	aa	
AA	$p^4 + p^3q = p^3$	$p^3q + p^2q^2 = p^2q$	0	p^2
Aa	$p^3q + p^2q^2 = p^2q$	$p^3q + 2p^2q^2 + pq^3 = pq$	$p^2q^2 + pq^3 = pq^2$	$2pq$
aa	0	$p^2q^2 + pq^3 = pq^2$	$pq^3 + q^4 = q^3$	q^2
Сумма	p^2	$2pq$	q^2	1

(в типе скрещивания он на первом месте). Аналогично, генотипы Aa и aa представлены строчками с 4-й по 6-ю и с 7-й по 9-ю. Складывая вероятности, соответствующие этим строчкам, найдем вероятности генотипов в потомстве, имеющих данный генотип отца (табл. 7).

Если A — доминантный ген, то, объединяя третий и четвертый столбцы табл. 6 и складывая соответствующие вероятности, получим таблицу вероятностей фенотипов в потомстве. Пользуясь ею и формулами Байеса, можно рассчитать вероятности типов скрещивания при заданном фенотипе потомства.

Пусть, например, родители и потомок имеют доминантный фенотип. Какова вероятность гетерозиготности обоих родителей?

Найдем сначала вероятность доминантного фенотипа потомка. Так как дополнительно известно, что оба родителя имеют доминантный фенотип, то нужно рассмотреть только скрещивания типов S_1 , S_2 , S_4 и S_5 . Иными словами, мы рассматриваем условную вероятность доминантного типа потомка при доминантном фенотипе родителей. Но для простоты будем писать кратко $P(D)$ вместо $P(D/S_1 + S_2 + S_4 + S_5)$. По формуле полной вероятности

$$\begin{aligned}
 P(D) &= P(AA + Aa) = P(S_1)P(D/S_1) + \\
 &+ P(S_2)P(D/S_2) + P(S_4)P(D/S_4) + P(S_5)P(D/S_5) = \\
 &= p^4 + 2p^3q + 2p^3q + 3p^2q^2 = \\
 &= p^2(p^2 + 2pq + q^2 + 2pq + 2q^2) = p^2(1 + 2q) = \\
 &= p^2(3 - 2p).
 \end{aligned}$$

Отсюда по формуле Байеса

$$P(S_5/D) = \frac{P(S_5) \cdot P(D/S_5)}{P(D)} = \frac{3p^2q^2}{p^2(3-2p)} = \frac{3q^2}{3-2p}.$$

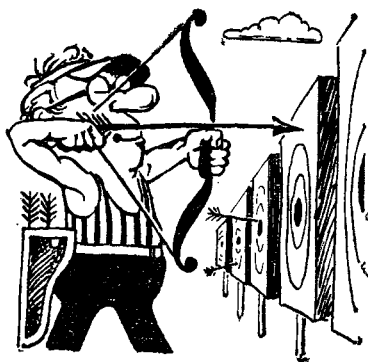
Аналогично вычисляются вероятности гетерозиготности одного из родителей:

$$P(S_2/D) = P(S_4/D) = \frac{2pq}{3-2p}.$$

Подобным образом находятся вероятности генотипов и фенотипов по признакам, сцепленным с полом. Возможные типы скрещивания при этом и условия вероятности генотипов потомства определяются по табл. 3. Подробности можно найти в книге Ч. Ли, на которую мы уже ссылались.

Дискретная случайная величина

1. Числовые события



Во всех сферах человеческой деятельности чаще других встречаются испытания, результатом которых является некоторое число — размер детали или выигрыша, отклонение от цели при стрельбе, число родившихся мальчиков, телефонных вызовов на АТС или купленных билетов за единицу времени, число особей в потомстве, обладающих определенным признаком, концентрация примесей, давление, температура, масса и т. д. и т. п. Даже в тех случаях, когда событие носит качественный характер (мальчик или девочка, герб или номинал, черный или белый и т. п.), мы можем свести дело к числам, приписав, например, одному исходу число 0, а другому 1. Более того, нас, как правило, интересует не только само событие, но и число таких событий в данной серии испытаний.

Понятно, что испытания с числовыми исходами привлекают особенное внимание исследователей. Их изучению и посвящено основное содержание теории вероятностей.

Случайное событие, состоящее в появлении некоторого числа, называется *числовым случайным событием*. Полная группа числовых событий называется *случайной величиной*, а сами эти события — *значениями случайной величины*.

Пример 1. При бросании игральной кости числовым случайным событием является число очков на верхней грани. Эти события образуют полную группу. Следовательно, их набор — случайная величина, а числа 1, 2, 3, 4, 5, 6 — значения этой случайной величины.

Пример 2. При бросании двух монет герб может выпасть 0, 1 и 2 раза. Набор этих числовых событий — случайная величина; 0, 1, 2 — ее значения.

В определении случайной величины, которое мы дали, предполагается, что эта величина принимает конечное число значений. В некоторых моделях приходится рассматривать случайную величину со счетным множеством значений: $x_1, x_2, \dots, x_n, \dots$. Случайные величины, принимающие конечное или счетное множество значений, называются *дискретными*.

Формально говоря, все случайные величины, встречающиеся на практике, принимают конечное или, в крайнем случае, счетное множество значений. Так, число мальчиков в тысяче новорожденных может быть равно любому целому числу от 0 до 1000; размер детали или плода измеряют целым числом, например, сантиметров или микрометров, уклонение при стрельбе — целым числом сантиметров или метров и т. д. Однако если значений достаточно много и они «плотно» заполняют некоторый промежуток, как микрометры в сантиметре, то при математическом моделировании удобнее считать, что значением случайной величины может быть любое число из данного промежутка, например, что длина детали может быть любым числом от 10 до 12 см. Подобные случайные величины называют *непрерывными*.

В этой главе мы остановимся на дискретных случайных величинах, а в следующей рассмотрим непрерывные.

2. Закон распределения

Указание множества всех возможных значений дискретной случайной величины недостаточно для ее характеристики. Разные случайные величины могут иметь одни и те же значения, но принимать их с разными вероятностями. Так будет, например, при бросании стандартной игральной кости и игральной кости со смещенным центром тяжести. Поэтому важнейшей характеристикой случайной величины является *закон распределения вероятностей*. Этот закон для дискретной случайной величины задают в виде таблицы, где в верхней строчке перечислены значения случайной величины X , а в нижней — вероятности P , с которыми

Таблица 8

X	x_1	x_2	\dots	x_n
P	p_1	p_2	\dots	p_n

Таблица 9

X	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

Таблица 10

X	0	1	2
P	1/4	1/2	1/4

Таблица 11

X	C
P	1

она их принимает (табл. 8). Например, если X — число выпавших очков при бросании стандартной игральной кости, то закон распределения задается табл. 9. Аналогично если X — число выпавших гербов при бросании двух монет, то закон распределения задается табл. 10.

Поскольку множество числовых событий, составляющих случайную величину, представляет собой полную группу событий, сумма вероятностей этих событий должна быть равна 1. Таким образом, для любой дискретной случайной величины X с конечным множеством значений x_1, x_2, \dots, x_n имеем

$$p_1 + p_2 + \dots + p_n = 1. \quad (1)$$

Если множество значений случайной величины счетно: $x_1, x_2, \dots, x_n, \dots$, то вместо выполнения равенства (1) требуется, чтобы ряд $\sum_{i=1}^{\infty} p_i$ сходил к его сумма была равна 1. Это требование, как и равенство (1), называют *условием полноты*.

В частном случае дискретная случайная величина может принимать одно-единственное значение, и веро-

ятность этого события, следовательно, равна 1. Такую случайную величину будем называть *постоянной*. Ее закон распределения очень прост (табл. 11).

3. Схема Бернулли.

Биномиальное распределение

Вывод старинной формулы. Предположим, что в одних и тех же условиях проводится n независимых испытаний, в каждом из которых событие A может появиться или не появиться. Пусть вероятность наступления события A в каждом испытании одна и та же и равна p . Следовательно, вероятность ненаступления равна $q = 1 - p$. Такая схема испытаний называется *схемой Бернулли*, по имени Якоба Бернулли, впервые рассмотревшего ее в своей книге «Искусство предположений».

Примерами испытаний по схеме Бернулли могут служить многократные подбрасывания монет и игральные кости, извлечения разноцветных шаров из одной и той же урны (с возвращением вынутого шара), стрельба по мишени из одного и того же орудия, контроль одинаковых изделий в партии и т. п.

Со схемой Бернулли связаны два типа случайных величин. Один из них — X_i — число наступлений события A в i -м испытании. Событие A может наступить 1 раз с вероятностью p и 0 раз с вероятностью q . Поэтому закон распределения X_i задается табл. 12.

Другая величина — X — число наступления события A во всей серии из n испытаний. Очевидно, X может принимать значения $0, 1, 2, \dots, k, \dots, n$. Найдем закон распределения X , т. е. подсчитаем вероятность того, что событие A наступит ровно k раз. Обозначим эту вероятность $p_n(k)$. Найдем сначала вероятность совмещения $B = \underbrace{A \cdot A \dots A}_{k \text{ раз}} \cdot \underbrace{\bar{A} \cdot \bar{A} \dots \bar{A}}_{n-k \text{ раз}}$. Так как

Т а б л и ц а 12

X_i	1	0
P	p	q

$P(A) = p, P(\bar{A}) = q$ и все события независимы, по формуле умножения

$$P(B) = p^k \cdot q^{n-k}.$$

Наступление совмещения B как раз и означает,

X	0	1	...	k	...	n
P	$C_n^0 q^n$	$C_n^1 p q^{n-1}$...	$C_n^k p^k q^{n-k}$...	$C_n^n p^n$

что событие A произошло ровно k раз. Но такое совмещение при k , отличном от 0 и n , не единственно. Например, при $k=3$ событие A может наступить в первом, втором и третьем испытаниях, но может — в первом, четвертом и седьмом или третьем, пятом и десятом. В общем случае совмещений типа B будет столько, сколько сочетаний по k испытаний можно образовать из n испытаний, т. е. C_n^k . Поэтому по формуле сложения

$$P_n(k) = C_n^k p^k q^{n-k}. \quad (2)$$

Равенство (2), называемое *формулой Бернулли*, дает закон распределения вероятностей наступления события A в серии из n испытаний (табл. 13). Этот закон называют еще *биномиальным*. Объясняется это тем, что $P_n(k) = C_n^k \cdot p^k \cdot q^{n-k}$ является общим членом в разложении бинома Ньютона

$$(q + p)^n = C_n^0 \cdot q^n + C_n^1 \cdot p q^{n-1} + \dots \\ \dots + C_n^k \cdot p^k \cdot q^{n-k} + \dots + C_n^n \cdot p^n.$$

Так как в нашем случае $p + q = 1$, из формулы бинома Ньютона получаем

$$\sum_{k=0}^n P_k(k) = 1,$$

как и положено для полной группы событий.

В 1988 году исполнилось 275 лет с тех пор, как формула Я. Бернулли была впервые опубликована — через 8 лет после смерти автора. Несмотря на свой более чем почтенный возраст, эта формула и связанная с ней схема Бернулли находят многочисленные приложения в самых различных областях современной науки. Приведем для иллюстрации несколько примеров.

Вероятность нескольких удач. Во многих реальных ситуациях наступление события A в испытании связывается с некоторым положительным эффектом — появлением исправной детали, поражением мишени при стрельбе, излечением больного и т. п. Поэтому наступление такого события называют успехом или удачей. При заданной вероятности одного успеха p мы по теореме умножения находили вероятность n успехов. Такая вероятность равна p^n . Однако задача при этом формулировалась очень жестко. Нужно было, чтобы успех наступил в каждом из n испытаний. В схеме Бернулли это только один из крайних вариантов. А вообще допускается, чтобы успех наступал не во всех, а лишь в k каких-то испытаниях, где $0 \leq k \leq n$. Вероятность наступления успеха в k испытаниях из n и вычисляется по формуле Бернулли.

Например, вероятность поражения мишени при одном выстреле равна 0,9. Какова вероятность того, что из 5 выстрелов ровно 3 будут удачными?

В данном случае $p = 0,9$, $q = 1 - 0,9 = 0,1$, $n = 5$, $k = 3$ и по формуле Бернулли мы можем написать

$$P_5(3) = C_5^3 \cdot 0,9^3 \cdot 0,1^2 = \frac{5!}{3!2!} \cdot \frac{9^3}{10^5} \approx 0,073.$$

Как видим, вероятность не очень велика. Объясняется это высокой вероятностью поражения мишени при каждом отдельном выстреле и, следовательно, малой вероятностью 2 промахов из 5 выстрелов.

Более разумно при большом p ставить другой вопрос: какова вероятность того, что из 5 выстрелов не менее 3 будут удачными? Эта вероятность выражается суммой $P_5(3) + P_5(4) + P_5(5)$. Вместо этой суммы удобнее найти равную ей разность $1 - [P_5(0) + P_5(1) + P_5(2)]$, так как в последнем случае не придется возводить 0,9 в четвертую и пятую степень. Имеем

$$P_5(0) = C_5^0 \cdot 0,9^0 \cdot 0,1^5 = 0,00001,$$

$$P_5(1) = C_5^1 \cdot 0,9^1 \cdot 0,1^4 = \frac{5 \cdot 9}{10^5} = 0,00045,$$

$$P_5(2) = C_5^2 \cdot 0,9^2 \cdot 0,1^3 = \frac{10 \cdot 9}{10^5} = 0,0081.$$

Малость найденных вероятностей также не удивительна. Если вероятность поражения при одном выстреле велика, то вероятность того, что 5 выстрелов подряд будут неудачными, должна быть мала. Складывая найденные вероятности и вычитая их сумму из единицы, найдем вероятность удачи не менее чем в 3 выстрелах из 5: она равна 0,99.

Рассмотрим задачу на эту же тему из области медицины. Применяемый метод лечения приводит к выздоровлению в 80 % случаев. Какова вероятность того, что из 5 больных выздоровят 4?

Казалось бы, поскольку 4 составляет 80 % от 5 и именно в 80 % гарантируется успех лечения, вероятность такого успеха должна быть равна 1. Однако это не так. В данном случае $p = 0,8$, $q = 0,2$, $n = 5$, $k = 4$. Поэтому по формуле Бернулли

$$P_5(4) = C_5^4 \cdot 0,8^4 \cdot 0,2^1 = \frac{5 \cdot 8^4 \cdot 2}{10^5} \approx 0,41.$$

Вероятность оказалась меньше половины.

Что произойдет с этой вероятностью, если увеличить эффективность лечения? Казалось бы, она должна вырасти. Однако при $p = 0,9$ получаем

$$P_5(4) = C_5^4 \cdot 0,9^4 \cdot 0,1^1 = \frac{5 \cdot 9^4}{10^5} = 0,33.$$

Вероятность уменьшилась! Эти парадоксальные на первый взгляд факты объясняются так же, как в задаче со стрельбой. При достаточно большой эффективности лечения выздоровление только 4 больных из 5 оказывается маловероятным событием.

Если сформулировать более естественную задачу и поинтересоваться вероятностью выздоровления не менее 4 больных из 5, то при $p = 0,8$ получим

$$P_5(4) + P_5(5) = 0,41 + 0,8^5 = 0,41 + 0,33 = 0,74,$$

а при $p = 0,9$ будем иметь

$$P_5(4) + P_5(5) = 0,33 + 0,9^5 = 0,33 + 0,59 = 0,92.$$

Из этого примера видно, что более или менее убедительные, но туманные доводы «здравого смысла» могут быть опровергнуты точным расчетом.

Рассмотрим, наконец, задачу из области генетики. В популяции плодовых мушек 25 % имеют мутацию глаз. Какова вероятность того, что из 10 взятых наугад мушек 6 имеют эту мутацию. По формулировке задача напоминает *задачу о выборе*. Но между этими задачами есть существенная разница. В задаче о наборе указываются общее число объектов N и число объектов, обладающих отмеченным свойством M . Если в рамках этой задачи отбирать объекты последовательно, один за другим, то вероятность появления объекта с отмеченным свойством не остается неизменной как в схеме Бернулли. Она зависит от того, какие объекты были отобраны ранее.

В сформулированной задаче о мушках предполагается, что популяция достаточно велика и извлечение для испытания 10 мушек практически не меняет процентного содержания мушек с мутацией, т. е. не меняет вероятности появления мушки с мутацией от испытания к испытанию. Поэтому для решения задачи можно применить формулу Бернулли. В данном случае $p = 0,25$ и

$$P_{10}(6) = C_{10}^6 \cdot p^6 \cdot q^4 = \frac{10!}{6!4!} \left(\frac{1}{4}\right)^6 \left(\frac{3}{4}\right)^4 = \frac{7 \cdot 8 \cdot 9 \cdot 10}{4!} \cdot \frac{81}{4^{10}} = 0,016.$$

Аналогично вычисляя $P_{10}(k)$, $0 \leq k \leq 10$, и складывая нужные вероятности, найдем вероятность того, что из 10 взятых наугад мушек отмеченной мутацией обладают более половины, менее двух, не менее одной и т. п.

Вероятность рождения потомков с заданными признаками.

Во многих биологических задачах схемы Бернулли серией испытаний является последовательность рождений, а событием A — рождение потомка с заданными признаками. Например, вероятность рождения мальчика $p = 1/2$. Какова вероятность того, что в семье с 3 детьми 2 мальчика и 1 девочка?

По формуле Бернулли

$$P_3(2) = C_3^2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right) = \frac{3}{8}.$$

Аналогично решаются задачи с другими признаками. Например, если родительская популяция состоит из гетерозигот Aa , то при скрещивании $Aa \times Aa$ с

вероятностью $1/4$ получается генотип AA , с вероятностью $1/2$ — генотип Aa и с вероятностью $1/4$ — генотип aa . При этом вероятность доминантного фенотипа, например вероятность кареглазого потомка, $p = P(AA + Aa) = 3/4$. Таким образом, вероятность того, что из 6 детей 4 будут кареглазыми, а 2 — голубоглазыми, вычисляется по формуле

$$P_6(4) = C_6^4 \cdot p^4 \cdot q^2 = \frac{6!}{4!2!} \left(\frac{3}{4}\right)^4 \left(\frac{1}{4}\right)^2 = 0,30.$$

Точно так же, рассмотрев популяцию с двумя признаками (см. табл. 4), по вероятности фенотипа потомка, например вероятности курчавого голубоглазого ($p = 3/16$), можно найти вероятность двух таких потомков из шести; трех из четырех и вообще k из n .

Задача об экстрасенсе. Обычный человек примерно в половине случаев правильно угадывает, в какой руке спрятан мелкий предмет. Предположим, что верный ответ получен в трех случаях из четырех. Случайно ли это? Или при таком результате можно говорить о необычных способностях угадывающего?

Если принять вероятность угадывания в норме $p = 1/2$, то по формуле Бернулли

$$P_4(3) = C_4^3 \cdot p^3 \cdot q = 4 \left(\frac{1}{2}\right)^3 \cdot \frac{1}{2} = \frac{1}{4}.$$

Как видим, вероятность не столь уж мала. Примерно каждый четвертый нормальный человек правильно угадает в трех случаях из четырех.

Но вот верный ответ получен в девяти случаях из десяти. Какова вероятность такого угадывания у нормального человека? По формуле Бернулли

$$P_{10}(9) = C_{10}^9 \left(\frac{1}{2}\right)^9 \cdot \frac{1}{2} = 10 \cdot \frac{1}{2^{10}} \simeq 0,01.$$

Таким образом, нормальный человек лишь в одном случае из 100 может случайно продемонстрировать такой результат. И если подобное угадывание происходит чаще, то можно, по-видимому, говорить, что угадыватель — экстрасенс (или мистификатор).

Задача о шахматном матче. Двум шахматистам равного класса предстоит играть матч между собой. Один из них — чемпион — сохранит свое звание,

если выиграет половину партий (ничьи в расчет не принимаются). Что вероятнее — выиграть две партии из четырех или три партии из шести?

На первый взгляд кажется, что, поскольку класс шахматистов примерно равный и каждый раз требуется выиграть половину партий, искомые вероятности должны быть равны между собой и равны $1/2$. Однако это не так.

В самом деле, не учитывая психологических моментов борьбы, можно считать, что вероятность выигрыша каждой партии, например, чемпионом $p = 1/2$ и вероятность певыигрыша (т. е. ничьих или проигрыша) $q = 1/2$. Тогда по формуле Бернулли

$$P_4(2) = C_4^2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{3}{8}, \text{ но } P_6(3) = C_6^3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^3 = \frac{5}{16}.$$

Как видно, вероятности $P_4(2)$ и $P_6(3)$ различны, причем $P_6(3)$ меньше. Подобным образом нетрудно найти вероятности

$$P_8(4) = \frac{35}{128}, P_{10}(5) = \frac{63}{256}, P_{12}(6) = \frac{231}{1024} \text{ и др.}$$

и заметить, что они убывают с ростом числа партий в матче.

Когда формула Бернулли оказывается бессильной. Во всех примерах, которые мы рассмотрели, n было не слишком велико, и мы без особого труда справились с вычислениями. При больших n , например при рассмотрении больших популяций, в которых исследованию нужно подвергнуть все особи, трудности вычисления сильно возрастают. Рассмотрим для примера партию деталей в 1000 штук. Предположим, что вероятность изготовления детали с дефектом равна $p = 0,005$. Какова вероятность того, что в этой партии 10 деталей окажется с дефектом?

По формуле Бернулли имеем

$$P_{1000}(10) = C_{1000}^{10} \cdot 0,005^{10} \cdot 0,995^{990} = \frac{1000! 5^{10} 995^{990}}{10! 990! 1000^{1000}}.$$

Ответ как будто получен. Но найти значение этой дроби совсем не просто. Попятно, что трудности еще более возрастут, если мы поставим задачу о вероятности появления не более 10 бракованных деталей,

не менее 10, но не более 20 и т. п. В таких ситуациях применяют методы, позволяющие приближенно вычислять $P_n(k)$. Точнее говоря, биномиальное распределение $P_n(k)$ при больших n приближенно заменяют другим распределением, более удобным для вычислений. С одним из таких распределений мы познакомимся в следующем разделе.

4. Распределение Пуассона

Замечательное приближение. Рассмотрим серию испытаний схемы Бернулли, в которых число исследуемых объектов очень велико, а вероятность наступления события A в отдельном испытании очень мала. Таким образом, если взять даже большую порцию испытаний, скажем n , то событие A в ней произойдет не более k раз, где k много меньше n . Такие события называются *редкими*.

Например, завод в течение месяца или года выпускает несколько тысяч деталей с малой вероятностью брака. В качестве порции n испытаний удобно взять 1000. В партии из 1000 деталей бракованная деталь может встретиться не более 10—15 раз, что много меньше 1000. Аналогично, в большой популяции рецессивный или мутагенный признак могут иметь 1—2 особи из нескольких сотен или тысяч особей. Таким образом, появление бракованной детали или особи с рецессивным или мутагенным признаком — редкое событие.

Часто в качестве порции испытуемых объектов удобно брать не какое-то их фиксированное число, например 1000, а единицу времени, в течение которой проводятся испытания, или единицу площади, на которой расположены испытуемые объекты. Примером может служить число остановок однотипных станков в течение смены в большом цехе. Станок случайно, т. е. в отсутствие общей аварии, может остановиться каждую минуту. Таким образом, испытанию подвергаются различные минуты рабочего времени. Число минут, отмеченных остановками станков, много меньше общего числа минут в смене. Поэтому остановки станков можно считать редкими событиями.

Аналогично рассуждают, когда число мерных квадратов, на которых обнаружено данное растение, много

меньше общего числа мерных квадратов на исследуемом участке растительной поверхности.

Другими примерами могут служить число неправильных соединений на АТС в течение часа; число опечаток на страницу; количество электронов, вылетающих с накаливаемого катода за минуту; количество радиационных импульсов, исходящих от единицы массы активированного вещества в минуту; число автомобильных катастроф в неделю; число травм на производстве в месяц; число мутаций в единицу времени или на число поколений; число поврежденных хромосом при облучении в данном объеме популяции; число ошибок при сборке молекул белков и нуклеиновых кислот в единицу времени и т. п.

Для описания вероятностей подобных случайных величин удобным оказывается закон распределения Пуассона:

$$P(k) = \frac{a^k}{k!} e^{-a}, \quad k = 0, 1, 2, \dots, \quad (3)$$

где a — некоторый параметр. Таким образом, случайная величина с законом распределения Пуассона принимает значения $0, 1, 2, \dots, k, \dots$ с вероятностями

$$P(0) = e^{-a}, \quad P(1) = e^{-a} \cdot a,$$

$$P(2) = \frac{e^{-a}}{2} a^2, \dots, \quad P(k) = \frac{e^{-a}}{k!} a^k.$$

Разумеется, на практике вероятности больших значений равны нулю. Например, вероятность того, что при нормальных условиях в течение часа откажут все 800 станков цеха, практически равна нулю. Однако в каждой конкретной ситуации свои пределы реальных значений k , а формула (3) при соответствующем подборе a может служить моделью всех подобных ситуаций.

Указав вероятность $P(k)$, мы тем самым указываем долю тех порций испытаний, в которых событие A произошло k раз. Например, если вероятность одной ошибки на страницу равна 0,1, а двух ошибок — 0,02, то это означает, что примерно 1 из 10 страниц содержит одну ошибку и 1 из 50 страниц — две ошибки.

Как же должна быть сформулирована конкретная задача, чтобы для ее решения можно было применить распределение Пуассона? Какой смысл имеет параметр a и как его подобрать? Чтобы ответить на эти вопросы, проще всего обратиться к первоисточнику. Впервые распределение Пуассона появилось в его статье «Исследования о вероятности приговоров в уголовных и гражданских делах», опубликованной в 1837 году. Пуассон доказал, что формулой (3) можно приближенно заменить биномиальное распределение $P_n(k)$, если n достаточно велико, а p мало.

Прежде чем доказывать теорему Пуассона, сделаем два замечания.

Сначала уточним понятие редкого события. В схеме Бернулли мы заранее фиксировали число испытаний n и вероятность p наступления события A в одном испытании. Теперь нам придется рассматривать серии с различными n . Так как при разных n вероятность p может быть разной, вместо p будем писать p_n , а вместо q — соответственно q_n . Положим далее $a_n = n \cdot p_n$. Смысл величины a_n станет ясным, если вспомнить о связи вероятности и относительной частоты. Например, если вероятность встретить дефектную деталь $p = 0,012$, то в партии из $n = 1000$ деталей примерно 12 будут дефектными. Среднее число дефектных деталей — 12 — это и есть $a_n = p \cdot n$. Так вот, в дальнейшем мы будем рассматривать серии испытаний, в которых $a_n \rightarrow a$ при $n \rightarrow \infty$, где a — некоторое число, отличное от нуля. Это означает, что с ростом длины серии n среднее число наступлений события A в серии стабилизируется, стремится к некоторому пределу a .

Второе замечание посит технический характер. В математическом анализе доказывается, что $\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = e$. Это так называемый второй замечательный предел. Отсюда следует, что

$$\lim_{n \rightarrow 0} \left(1 + \frac{m}{n}\right)^n = \lim_{n \rightarrow 0} \left(1 + \frac{m}{n}\right)^{\frac{n}{m} \cdot m} = e^m.$$

Теперь все готово, чтобы сформулировать и доказать теорему, отметившую недавно свой 150-летний юбилей.

Теорема Пуассона. Пусть p_n — вероятность наступления события A в каждом испытании серии из n испытаний схемы Бернулли. Тогда, если $a_n = n \cdot p_n \rightarrow a$ при $n \rightarrow \infty$, $0 < a < \infty$, то биномиальное распределение $P_n(k)$ при $n \rightarrow \infty$ стремится к распределению Пуассона с параметром a .

Иными словами, если n достаточно велико, то для каждого $k = 0, 1, 2, \dots$ вероятность

$$P_n(k) = C_n^k \cdot p_n^k \cdot q_n^{n-k} = \frac{a_n^k}{k!} e^{-a_n}.$$

В самом деле,

$$\begin{aligned} P_n(k) &= C_n^k \cdot p_n^k \cdot q_n^{n-k} = \frac{n(n-1) \dots (n-k+1)}{k!} \times \\ &\times \left(\frac{a_n}{n}\right)^k \left(1 - \frac{a_n}{n}\right)^{n-k} = \frac{a_n^k}{k!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \dots \frac{(n-k+1)}{n} \times \\ &\times \left(1 - \frac{a_n}{n}\right)^{n-k} \cdot \left(1 - \frac{a_n}{n}\right)^k. \end{aligned}$$

Переходя в этом равенстве к пределу при $n \rightarrow \infty$, заметим, что первый сомножитель стремится к $\frac{a^k}{k!}$, все остальные, кроме последнего, — к 1, а последний сомножитель в силу второго замечательного предела стремится к e^{-a} . Это и доказывает теорему.

На практике число испытаний n всегда конечно и трудно проверить, выполняется ли условие $n \cdot p_n \rightarrow a$ при $n \rightarrow \infty$. Однако из этого условия следует, что $p_n = \frac{a_n}{n} \rightarrow 0$. Поэтому если в схеме Бернулли n вели-

ко, а p мало, то в качестве значения параметра a берут $a = n \cdot p$ и биномиальное распределение $P_n(k)$ заменяют распределением Пуассона с этим значением a .

Редкие болезни, редкие признаки. Многие болезни, к счастью, достаточно редки или становятся таковыми после принятия профилактических и лечебных мер. Однако даже при самых благоприятных условиях в больших популяциях все же встречается некоторое число больных редкими болезнями. Распределение Пуассона дает вероятность таких издержек в нормальной ситуации. Если в наблюдаемой популяции, например в конкретном городе, больные встречаются чаще, чем это предписывается законом

Пуассона для всей страны, то это говорит о нарушении нормальных условий в данном городе, о необходимости выяснения причин, принятия срочных мер и т. д.

Например, при введении вакцины против полиомиелита иммунитет создается в 99,99 % случаев. Какова вероятность того, что из 10 000 вакцинированных детей заболит 1?

Вероятность заболеть $p = 0,0001$, число испытаний $n = 10\,000$. Поэтому $a = n \cdot p = 1$, и по формуле Пуассона имеем

$$P(1) \simeq \frac{1}{1!} e^{-1} = 0,368.$$

Аналогично, вероятность, что заболеют 2 ребенка:

$$P(2) \simeq \frac{1^2}{2!} e^{-1} \approx 0,184,$$

а вероятности заболевания 3 и 4 детей соответственно равны

$$P(3) \simeq \frac{e^{-1}}{3!} = 0,061, \quad P(4) \simeq \frac{e^{-1}}{4!} \simeq 0,015.$$

Добавим сюда вероятность того, что ни один ребенок не заболел $P(0) = P(1) = 0,368$, и вероятность того, что заболели 5 детей $P(5) = e^{-1} \cdot 5! = 0,001$.

Если приять, что 10 000 новорожденных — это годовая норма, скажем, крупного районного роддома, то примерно в 73 % таких домов полиомиелитом заболевает не более одного ребенка в год; в 18 % — два ребенка в год; в 6 % — три и в 1,5 % — четыре ребенка в год. Если же в каком-то роддоме заболело 5 детей — то это чрезвычайное происшествие. Вероятность такого события равна 0,001.

Аналогичные расчеты можно провести по детской смертности, туберкулезу, врожденным аномалиям и признакам. Например, 1 из 600 детей рождается с болезнью Дауна. Следовательно, в каждом микрорайоне, где проживает 3000 детей, в среднем будет $a = np = 3000 \cdot \frac{1}{600} = 5$ детей, страдающих такой болезнью.

При этом вероятность рождения 10 детей с синдромом Дауна в микрорайоне равна

$$P(10) = \frac{a^{10}}{10!} e^{-a} = \frac{5^{10}}{10!} e^{-5} = 0,018.$$

Таким образом, подобных микрорайонов должно быть не более 2 из 100.

Часто задаются не вероятностью p и число испытаний n , а сразу характерное значение параметра $a = n \cdot p$, т. е. среднее значение наступления события A во всех больших сериях. Это значение a находят из эксперимента. Например, результаты наблюдений показали, что в травматологическое отделение в течение каждого часа дневного времени привозят примерно 3 пациентов. Какова вероятность того, что их будет только один или ни одного? В этом случае $a = 3$ и

$$P(1) = \frac{a}{1!} e^{-a} = 3e^{-3} = 0,15, \quad \text{а} \quad P(0) = \frac{a^0}{0!} e^{-a} = 0,05.$$

Таким образом, вероятность того, что в течение часа потребуется по крайней мере 1 хирург, равна $1 - P(0) = 0,95$, а что таких хирургов нужно будет не менее 2, равна $1 - P(0) - P(1) = 0,8$. Как видно, отделение травматологии нельзя закрывать на обед.

Ошибки, сбои, опечатки. Распределение Пуассона успешно применяется для описания редких отклонений от нормы в хорошо отлаженном, например автоматическом, процессе. Предположим, что на АТС в течение часа происходит примерно 10 неправильных соединений. Взяв это значение в качестве a , по формуле Пуассона $P(k) = \frac{a^k}{k!} e^{-a}$ найдем вероятность

только одного неправильного соединения, только двух, только трех и т. д. В частности, можно найти и вероятность безошибочной работы в течение часа $P(0) = e^{-10}$. Она исчезающе мала.

Аналогично если при автоматическом наборе печатного текста среднее число опечаток, приходящееся на страницу, равно 0,5, то доля страниц с одной опечаткой $P(1) = \frac{0,5}{1!} e^{-0,5} = 0,3$, доля страниц с двумя опечатками $P(2) = \frac{0,5^2}{2!} e^{-0,5} = 0,075$, наконец, доля безошибочных страниц $P(0) = e^{-0,5} = 0,6$.

Подобный «автоматический набор» текста с ошибками происходит при «сборке» молекул белков и нуклеиновых кислот. Но в отличие от печатного текста ошибки в этом случае как раз интересны, так как ведут к появлению мутантов.

Если p — вероятность ошибки и n — длина слова, то вероятность получить слово с k ошибками вычисляется по формуле Бернулли

$$P_n(k) = C_n^k \cdot p^k \cdot q^{n-k},$$

Формально говоря, поскольку каждый раз приходится выбирать одну букву из алфавита мощностью m , вероятность ошибки, т. е. вероятность выбора неправильной буквы, $p = \frac{m-1}{m}$. Для белков $m = 20$, для нуклеиновых кислот $m = 4$. Однако выбор букв диктуется не столько комбинаторикой, сколько физико-химическими свойствами букв-мономеров. В идеальном случае сборка вообще могла быть безошибочной. Но из-за энергетических колебаний с небольшими вероятностями возникают сбои, ошибки. В обычных условиях вероятность ошибки $p = 0,01$. В эксперименте p можно уменьшить до 10^{-4} — 10^{-6} . При таких p и больших n вместо формулы Бернулли естественно воспользоваться приближением Пуассона

$$P_n(k) = P(k) = \frac{a^k}{k!} e^{-a}, \quad a = np,$$

В частности, вероятность безошибочной сборки $P_n(0) = q^n \approx e^{-a}$.

Характер мутанта существенно зависит от того, сколько и в каких местах произошли ошибки при сборке. Если указано место в слове, где может произойти ошибка, то, поскольку вместо правильной буквы можно поставить $m - 1$ ошибочных букв, всего таких неправильных слов с одной ошибкой в указанном месте можно собрать $m - 1$. Если место ошибки не фиксировать, т. е. допустить, что ошибка происходит в любом из n мест слова, то общее количество слов с одной ошибкой будет равно $N_1 = n(m - 1)$. Аналогично если зафиксировать k мест в слове, в которых может произойти ошибка, то всего можно собрать

$(m-1)^k$ таких неправильных слов. Если же k мест не фиксировать и рассматривать всевозможные сочетания k мест из n , то общее количество слов с k ошибками будет $N_k = C_n^k (m-1)^k$.

Будем считать, что все неправильные слова с k ошибками равновероятны. Тогда, чтобы получить вероятность образования фиксированного слова с k ошибками, нужно вероятность образования хотя бы одного из таких слов (вероятность суммы событий) $P_n(k)$ разделить на общее число таких слов N_k . Следовательно, вероятность появления определенного мутанта можно вычислить по формуле

$$P_k = \frac{P_n(k)}{N_k} = \frac{C_n^k \cdot p^k \cdot q^{n-k}}{C_n^k (m-1)^k} = \frac{p^k \cdot q^{n-k}}{(m-1)^k}.$$

Если n велико, а k мало, то эта дробь вычисляется с трудом. Ее можно оценить, воспользовавшись приближением Пуассона. Как мы отметили, $q^n \approx e^{-a}$. Поэтому

$$P_k = \frac{p^k \cdot q^n \cdot q^{-k}}{(m-1)^k} \approx e^{-a} \left(\frac{p \cdot q^{-1}}{m-1} \right)^k = e^{-a} \left(\frac{q^{-1} - 1}{m-1} \right)^k.$$

При анализе динамики числа мутантов ключевую роль играют скорости их образования. Эти скорости пропорциональны найденным вероятностям.

Закончим примером с дефектными деталями, который мы не смогли решить с помощью формулы Бернулли. Так как $p = 0,005$, $n = 1000$, $a = np = 5$, имеем

$$P_{1000}(10) \approx P(10) = \frac{5^{10} \cdot e^{-5}}{10!} = 0,018.$$

Сколько клеток, сколько лет. Мы уже не раз решали своего рода обратные задачи типа задачи о дублерах, в которых по заданной вероятности нужно определить значение параметра, обеспечивающее эту вероятность. Подобные задачи возникают и в связи с распределением Пуассона

$$P(k) = \frac{a^k}{k!} e^{-a}, \quad a = np.$$

Искомый параметром при заданных k , p и $P(k) = P$ является a или чаще $n = a/p$.

В общем виде при произвольном k эта задача далеко не простая, так как уравнение

$$\frac{a^k}{k!} e^{-a} = P \quad (4)$$

относительно a в конечном виде, т. е. с помощью элементарных функций, не решается. Однако в некоторых частных случаях можно без особого труда добиться успеха. Например, при $k=0$ получаем $e^{-a} = P$. Отсюда $-a = \ln P$ и, следовательно,

$$n = -\frac{\ln P}{p}. \quad (5)$$

Таким образом, если заданы вероятность p наступления редкого события в одном испытании и вероятность $P = P(0)$ ненаступления события в серии из n испытаний, то по формуле (5) можно найти число испытаний n .

Этот рецепт используется при оценке числа n объектов — молекул, клеток, растений, животных и т. п., более или менее равномерно распределенных на некотором участке. Для этого участок с объектами разбивается на ячейки, например квадраты. Пусть общее число ячеек равно N . В качестве события A будем рассматривать случайное попадание объекта в определенную ячейку. Очевидно, вероятность такого попадания $p = P(A) = 1/N$. Ячейки, вообще говоря, по-разному заполнены объектами. В некоторых из них, например, 5—6 объектов, в других — 3 или 2, а есть и такие, в которых нет ни одного.

Пусть m — общее число пустых ячеек. Тогда отношение m/N — доля пустых ячеек — равно вероятности того, что ни один объект из n не попал в определенную ячейку. Таким образом, $P = P(0) = m/N$. Подставив значения p и P в равенство (5), получим

$$n = -N \ln \frac{m}{N} = N \ln \frac{N}{m}. \quad (6)$$

На практике число пустых ячеек m обычно найти трудно. Это могут быть пустые квадратики на разграфленном лабораторном стекле с клетками крови, пустые ячейки в решетке с уснувшими плодовыми мушками и т. п. Часто в качестве m берут среднее

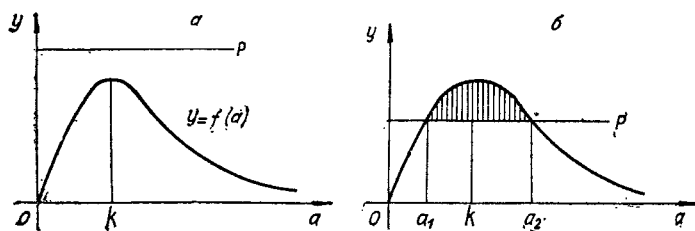


Рис. 12.

арифметическое нескольких экспериментов. Число N , естественно, известно. Зная m и N , по формуле (6) находят n .

Полезные выводы из уравнения (4) можно получить и в общем случае при $k \neq 0$. Исследуем для этого график функции $f(a) = \frac{a^k}{k!} e^{-a}$ при фиксированном k в области, где $a \geq 0$. Имеем, очевидно, $f(0) = 0$; $f(a) \rightarrow 0$ при $a \rightarrow \infty$; $f'(a) = \frac{a^{k-1} e^{-a}}{k!} (k - a)$. Отсюда следует, что при $k \neq 0$ график функции $y = f(a)$, $a \geq 0$, представляет собой импульсообразную кривую с максимумом в точке $a = k$, равным $\frac{k^k}{k!} e^{-k}$ (рис. 12).

Последнее обстоятельство ограничивает возможные значения P в уравнении (4). Если P задать так, что $P > f(k) = \frac{k^k}{k!} e^{-k}$, то не найдется ни одного значения a , для которого бы выполнялось равенство (4) (см. рис. 12, а). Например, при $k = 2$ значение P должно быть не больше $\frac{2^2}{2!} e^{-2} = 0,272$, а при $k = 3$ — не больше $\frac{3^3}{3!} e^{-3} = 0,226$ и т. д.

Если же P задать так, что $P < f(k) = \frac{k^k}{k!} e^{-k}$, то найдется некоторый промежуток значений a , $a_1 \leq a \leq a_2$, для которых выполняется неравенство

$$P(k) = \frac{a^k}{k!} e^{-a} \geq P \quad (7)$$

(см. рис. 12, б). Этот промежуток для каждого фиксированного k можно найти тем или иным приближенным методом.

Распределение Пуассона и, следовательно, функция $f(a)$ в настоящее время изучены очень детально. Составлены подробные таблицы вероятностей $P(k) = \frac{a^k}{k!} e^{-a}$ при различных значениях a . Эти вероятности обозначаются обычно $p_k(a)$ или $p_a(k)$. Пользуясь такими таблицами, можно по заданному P найти промежуток значений a , для которых выполняется неравенство (7).

Для примера рассмотрим задачу из теории эволюции. Известно, что вероятность появления мутантного гена равна $p = 10^{-6}$. Для того чтобы этот ген закрепился в популяции, нужно, чтобы ежегодно появлялось не менее 10 указанных мутаций. Точнее говоря, вероятность появления 10 мутаций в год должна быть достаточно большой. Каково максимальное значение P_0 этой вероятности? Каким должен быть размер популяции n , чтобы вероятность появления 10 мутантов в год была близка к P_0 ?

При $k = 10$ по таблице распределения Пуассона находим

$$P_0 = \max f(a) = f(10) = P_{10}(10) = \frac{10^{10}}{10!} e^{-10} \simeq 0,125.$$

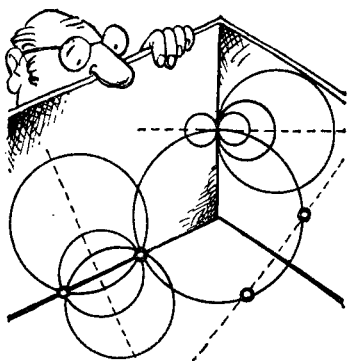
Это достаточно высокая вероятность, и она означает, что примерно раз в 8 лет появляются 10 мутантов. Пусть эта вероятность признана достаточной для закрепления мутантного гена. Мы получим близкую к P_0 вероятность, если $a \approx 10$, т. е. при $np \approx 10$. Отсюда

$$n \approx \frac{10}{p} = \frac{10}{10^{-6}} = 10^7.$$

Таким образом, в популяции, содержащей 10^7 особей, с вероятностью, примерно равной 0,125, ежегодно возникают 10 мутантов.

Если порог вероятности понизить, т. е. взять $P < P_0$, то расширится промежуток значений a . Он будет содержать, в частности, значения $a < 10$. Этим значениям будут соответствовать размеры популяции $n = a/p < 10^7$. Однако в таких меньших по размеру популяциях реже будут встречаться годы, в которых появляются 10 мутантов.

Действия со случайными величинами



На практике часто приходится рассматривать не одну, а сразу несколько случайных величин. Например, если стержень состоит из двух частей, каждая из которых изготавливается на отдельном станке, то длина стержня есть случайная величина, значения которой определяются двумя различными случайными величинами — размерами отдельных частей. Аналогично площадь прямоугольника определяется длинами его сторон, путь — скоростью и временем движения и т. д.

Из приведенных примеров видно, что нам необходимо ввести простейшие операции над случайными величинами — сумму и произведение. Но прежде нужно договориться о том, какие случайные величины мы будем считать равными.

1. Равенство двух случайных величин

Как и всякое множество, одна и та же случайная величина может быть описана различными способами. Отличаясь внешне, эти способы задают одно и то же множество числовых событий и вероятности их наступления. Поэтому случайная величина X со значениями x_i и вероятностями p_i , $i = 1, 2, \dots, n$, совпадает со случайной величиной Y со значениями y_k и вероятностями g_k , $k = 1, 2, \dots, m$, если $n = m$, $x_i = y_i$, $p_i = g_i$ и события случайной величины X наступают тогда и только тогда, когда наступают события случайной величины Y .

Последнее условие — одновременное наступление событий X и Y — очень важно. Чтобы это прояснить,

Таблица 14

X	1	0
P	1/2	1/2

Таблица 15

Y	1	0
P	1/2	1/2

рассмотрим два примера. Предположим, что независимо одна от другой бросаются две монеты и X — число выпадений герба на первой, а Y — на второй. Очевидно, что обе величины имеют *один и тот же* закон распределения (табл. 14, 15). Однако это *различные* случайные величины, так как события одной никак не связаны с событиями другой.

Рассмотрим теперь одну монету. Сторону, на которой изображен герб, пометим красной краской, а сторону с номиналом — белой. Пусть X — число выпадений герба при бросании этой монеты, а Y — число выпадений красного цвета. Тогда очевидно, что X и Y совпадают. Их описания только внешне отличаются одно от другого. Каждый раз, когда происходит событие X (выпадение герба), происходит и событие Y (выпадение красного цвета), и наоборот. Таким образом, X и Y представляют собой одно и то же множество событий и их вероятностей.

Если две случайные величины X и Y совпадают, то будем говорить, что они *равны*, и писать, как для чисел, $X = Y$.

2. Сумма случайных величин

Пусть имеются случайная величина X со значениями x_i и вероятностями p_i , $i = 1, 2, \dots, n$, а также случайная величина Y со значениями y_k и вероятностями g_k , $k = 1, 2, \dots, m$. Обозначим A_i событие « X принимает значение x_i » (кратко $X = x_i$), B_k — событие « Y принимает значение y_k » (кратко $Y = y_k$) и пусть $p_{ik} = P(A_i B_k)$ — вероятность совмещения $A_i B_k$.

Определение. Суммой двух случайных величин X и Y называются случайная величина $X + Y$, значения которой равны всевозможным суммам $x_i + y_k$, а соответствующие вероятности — вероятностям совмещений p_{ik} .

$X + Y$	$1 + 1 = 2$	$1 + 0 = 0 + 1 = 1$	$0 + 0 = 0$
P	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

Если две различные суммы значений совпадают, то их общему значению приписывается вероятность, равная сумме соответствующих вероятностей совмещений. Например, если $x_1 + y_2 = x_3 + y_4$, то значению $x_1 + y_2$ приписывается вероятность $p_{12} + p_{34}$.

Это определение вполне согласуется с нашим интуитивным представлением о сумме: чтобы составить стержень длиной $x_i + y_h$, мы должны иметь совмещение событий A_i и B_h .

Если события одной случайной величины не зависят от событий другой, то такие случайные величины будем называть *независимыми*. Для независимых величин по формуле умножения $p_{ih} = p_i p_h$.

Пример. Пусть X — число выпадений герба при бросании первой монеты, а Y — при бросании второй. Очевидно, X и Y — независимые случайные величины. Их распределения указаны в табл. 14 и 15. По определению, их сумма $X + Y$ имеет распределение, приведенное в табл. 16.

Как видно, случайная величина $X + Y$ — это число выпадений гербов при одновременном бросании двух монет.

В частном случае одно из слагаемых может быть постоянной величиной с законом распределения $X \equiv C$, $p = 1$. Обозначим постоянную случайную величину, принимающую значение C , той же буквой и рассмотрим закон распределения суммы $X + C$. По определению, значения этой суммы равны $x_1 + C$, $x_2 + C$, ..., ..., $x_n + C$, а соответствующие вероятности $p_1 \cdot 1 = p_1$, $p_2 \cdot 1 = p_2$, ..., $p_n \cdot 1 = p_n$. Таким образом, сумма $X + C$ принимает значения $x_i + C$ с вероятностями p_i .

Понятие суммы естественно обобщается на любое число случайных величин.

Пример. Пусть X и X_i — случайные величины схемы Бернулли. Тогда нетрудно установить, что

$$X = X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_i,$$

т. е. число появлений события A во всей серии испытаний равно сумме появлений этого события в i -х испытаниях.

В самом деле, ограничимся случаем $n = 3$. По формуле Бернулли случайная величина X принимает значения $0, 1, 2, 3$ с вероятностями $P_3(0) = q^3$, $P_3(1) = 3pq^2$, $P_3(2) = 3p^2q$ и $P_3(3) = p^3$. Рассмотрим теперь сумму $X_1 + X_2 + X_3$. По определению суммы, эта случайная величина принимает значение $0 = 0 + 0 + 0$ с вероятностью $q \cdot q \cdot q = q^3$, значение $1 = 1 + 0 + 0 = 0 + 1 + 0 = 0 + 0 + 1$ с вероятностью $pqq + qpq + qqp = 3pq^2$, значение $2 = 1 + 1 + 0 = 1 + 0 + 1 = 0 + 1 + 1$ с вероятностью $prq + pqr + qrp = 3p^2q$ и, наконец, значение $3 = 1 + 1 + 1$ с вероятностью $ppr = p^3$.

Таким образом, случайные величины X и $X_1 + X_2 + X_3$ принимают одинаковые значения одновременно и с одинаковыми вероятностями. Это означает, что они совпадают: $X = X_1 + X_2 + X_3$.

Для произвольного n доказательство аналогично.

3. Произведение случайных величин

Пусть, как и ранее, A_i означает событие $X = x_i$, B_k — событие $Y = y_k$, а $p_{ik} = P(A_i B_k)$ — вероятность совмещения $A_i B_k$.

О п р е д е л е н и е. Произведением двух случайных величин X и Y называется случайная величина, которая обозначается XY , значения которой равны всевозможным произведениям $x_i y_k$, а соответствующие вероятности — вероятностям совмещений p_{ik} .

Если два различных произведения значений совпадают, то их общему значению приписывается вероятность, равная сумме соответствующих вероятностей совмещений. Например, если $x_1 y_2 = x_3 y_4$, то значению $x_1 y_2$ приписывается вероятность $p_{12} + p_{34}$.

Это определение произведения случайных величин также хорошо согласуется с опытом.

П р и м е р. Пусть X — число выпадений герба на первой монете, а Y — число выпадений герба на второй. Тогда закон распределения вероятностей произведения XY имеет вид, указанный в табл. 17. Видно, что XY — это число выпадений сразу двух гербов при бросании двух монет.

XY	$1 = 1 \cdot 1$	$0 = 1 \cdot 0 = 0 \cdot 1 = 0 \cdot 0$
P	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$

В частном случае один из сомножителей может быть постоянной случайной величиной C . По определению, произведение XC принимает значения $x_1 \cdot C, x_2 \cdot C, \dots, x_n \cdot C$ с вероятностями $p_1 \cdot 1, p_2 \cdot 1, \dots, p_n \cdot 1$. Таким образом, произведение XC принимает значения $x_i C$ с вероятностями p_i .

В другом частном случае может быть $Y = X$. Тогда $XY = XX = X^2$. Найдем закон распределения X^2 . В этом случае $A_k = B_k$, и потому совмещение $A_i B_k = A_i A_k$ — невозможное событие, если $i \neq k$ (X не может одновременно принять два различных значения: x_i и x_k). Это значит, что все вероятности $p_{ik} = P(A_i A_k) = 0$, если $i \neq k$. Остаются только значения $x_i x_i = x_i^2$, которые приписываются с вероятностями $p_{ii} = P(A_i A_i) = P(A_i) = p_i$. Таким образом, если X принимает значения x_i с вероятностями p_i , то X^2 принимает значения x_i^2 с теми же вероятностями p_i . Это хорошо согласуется с опытом: круги площадью πR^2 встречаются так же часто, как радиусы R .

Понятие произведения естественно обобщается на любое число случайных величин.

Введенные операции суммы и произведения случайных величин подчиняются обычным законам:

$$1^\circ. X + Y = Y + X; XY = YX.$$

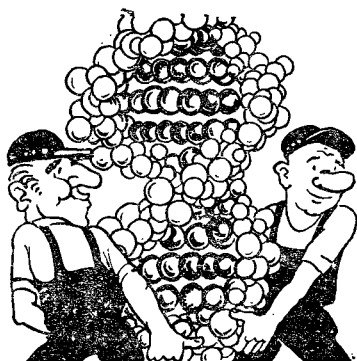
$$2^\circ. (X + Y) + Z = X + Y + Z; (XY)Z = XYZ.$$

$$3^\circ. (X + Y)Z = XZ + YZ.$$

Доказательства этих тождеств следуют непосредственно из определений. Следствием тождества 3° является формула сокращенного умножения:

$$(X + Y)^2 = X^2 + 2XY + Y^2.$$

Числовые характеристики дискретной случайной величины



Закон распределения вероятностей случайной величины исчерпывающе описывает множество ее значений и их вероятностей. Однако если значений много, таблица закона распределения становится громоздкой, труднообозримой. Возникает необходимость каких-то обобщений, компактных числовых характеристик случайной величины. Такими характеристиками являются *математическое ожидание* и *дисперсия*. Первая из них характеризует некоторый средний уровень значений случайной величины, а вторая — разброс, рассеяние значений от этого среднего уровня. Числовые характеристики случайной величины называют также ее *параметрами*.

1. Математическое ожидание

Будем рассматривать дискретную случайную величину X со значениями x_i и вероятностями p_i .

Определение. Математическим ожиданием дискретной случайной величины X называется сумма произведений ее значений на их вероятности. Обозначается математическое ожидание $M(X)$, или MX .

Таким образом,

$$M(X) = x_1 p_1 + x_2 p_2 + \dots + x_n p_n. \quad (1)$$

Пример. Пусть X — число выпадений герба при бросании двух монет. Так как свои значения 0, 1, 2 эта величина принимает с вероятностями $1/4$, $1/2$, $1/4$, то $M(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$.

Подчеркнем, что математическое ожидание $M(X)$ — уже не случайная величина, а некоторая постоянная, характеризующая закон распределения данной случайной величины X .

Выясним вероятностный смысл математического ожидания. Предположим, что на станке изготовлено n одинаковых деталей. Хотя станок и отрегулирован, размеры деталей не постоянны; например, их длина принимает различные значения из какого-то, пусть достаточно малого, промежутка. Предположим, что m_1 деталей имеют длину x_1 , m_2 — длину x_2 , ..., m_k — длину x_k . Тогда средняя арифметическая длина

$$\bar{x} = \frac{x_1 m_1 + x_2 m_2 + \dots + x_k m_k}{n}$$

или
$$\bar{x} = x_1 \cdot \frac{m_1}{n} + x_2 \cdot \frac{m_2}{n} + \dots + x_k \cdot \frac{m_k}{n}.$$

Как мы знаем, если n достаточно велико, то относительная частота $\frac{m_i}{n}$ колеблется около вероятности p_i появления детали длиной x_i . Таким образом, $\bar{x} \simeq \sum_{i=1}^k x_i p_i = M(X)$, т. е. математическое ожидание случайной величины — это «среднее арифметическое» всех ее значений (с учетом частоты их появлений). Поэтому математическое ожидание называют также *средним значением* случайной величины.

Математическое ожидание оказывается очень полезным во многих не только теоретических, но и сугубо практических вопросах. Если известно среднее количество материала, необходимого для изготовления одной детали, то легко найти количество материала, которое потребуется для партии из 1000 штук. Аналогично определяются количество рабочих мест в ремонтной мастерской, количество магазинов в микрорайоне, количество коек в больнице и т. п.

Приведем два примера из производственных областей. На студенческом вечере организуется лотерея, в которой разыгрывается 10 выигрышей ценой по 3 рубля, 20 выигрышей по 1 рублю и 80 выигрышей по 50 копеек. Предполагается продать 300 билетов. Каково математическое ожидание размера выигрыша для человека, купившего один билет?

Размер выигрыша за один билет — случайная величина. Обозначим ее X . Очевидно, X может принимать значения 3 рубля, 1 рубль, 0,5 рублей и 0 рублей. Нетрудно найти вероятности этих значений. По классической схеме находим

$$P(3) = \frac{10}{300}, \quad P(1) = \frac{20}{300}, \quad P(0,5) = \frac{80}{300}, \quad P(0) = \frac{300 - (10 + 20 + 80)}{300}.$$

Таким образом, закон распределения X может быть представлен табл. 18.

Отсюда легко найти математическое ожидание

$$M(X) = 3 \cdot \frac{1}{30} + 1 \cdot \frac{2}{30} + 0,5 \cdot \frac{8}{30} = \frac{9}{30} = 0,3.$$

Следовательно, после проведения лотереи каждый купивший 1 билет выиграет в «среднем» по 0,3 рубля. Понятно, что если этот «средний» выигрыш умножить на число участников лотереи, то получим $0,3 \cdot 300 = 90$ рублей. Это сумма, которую оргкомитет затратил на приобретение призов.

Т а б л и ц а 18

X	3	1	0,5	0
P	1/30	2/30	8/30	19/30

Т а б л и ц а 19

X	100	—900
P	0,927	0,073

Однако обычно проведение вечера связано с некоторыми организационными расходами — оплатой труда гардеробщицы, украшением зала и т. п. Чтобы получить от лотереи накопление для этих целей, можно пойти двумя путями: а) увеличить цену билетов; б) продать больше чем 300 билетов. В первом случае математическое ожидание не меняется и, следовательно, не убывает надежда на выигрыш каждого участника. Если требуется получить от лотереи 120 рублей, то цена одного билета должна быть $120/300 = 0,4$ рубля. Во втором случае не меняется цена билета — 30 копеек, и те же 120 рублей будут получены, если будет продано $120/0,3 = 400$ билетов. При этом заметно убы-

вает «средний» выигрыш. Легко найдем, что в этом случае

$$M(X) = 3 \cdot \frac{1}{40} + 2 \cdot \frac{1}{40} + 0,5 \cdot \frac{8}{40} = 0,225.$$

Можно выбрать и смешанную стратегию — продать, например, 350 билетов по 35 копеек и т. д.

Подобные соображения лежат в основе всех лотерей.

Рассмотрим еще пример. Демографами в результате специальных обследований и расчетов составляются так называемые *таблицы смертности*. Согласно этим таблицам, в СССР, например, вероятность того, что 40-летний человек доживет до 50 лет, равна 0,927 (таблицы 60-х годов). Если такой человек будет застрахован на 10 лет на сумму 1000 рублей со страховым взносом 100 рублей, то Госстрах получит доход 100 рублей с вероятностью 0,927 и вынужден будет выплачивать 1000 рублей с вероятностью $1 - 0,927 = 0,073$. Таким образом, доход Госстраха — это случайная величина с законом распределения, представленным табл. 19. В последней колонке стоит -900, а не -1000, так как из суммы выплаты удерживается страховая взнос.

Зная закон распределения, легко найдем средний доход Госстраха:

$$M(X) = 100 \cdot 0,927 - 900 \cdot 0,073 = 92,7 - 65,7 = 27.$$

Таким образом, каждое такое страхование приносит Госстраху «в среднем» 3%-ный годовой доход.

2. Свойства математического ожидания

Многочисленные приложения математического ожидания опираются на следующие его свойства.

1°. *Математическое ожидание постоянной равно самой этой постоянной:*

$$M(C) = C.$$

В самом деле, из закона распределения постоянной следует, что $M(C) = C \cdot 1 = C$.

2°. Постоянный множитель можно выносить за знак математического ожидания:

$$M(CX) = CM(X).$$

Действительно, из закона распределения CX следует, что

$$\begin{aligned} M(CX) &= Cx_1p_1 + Cx_2p_2 + \dots + Cx_np_n = \\ &= C \sum_{i=1}^n x_i p_i = CM(X). \end{aligned}$$

3°. Математическое ожидание суммы случайных величин равно сумме их математических ожиданий:

$$M(X + Y) = M(X) + M(Y).$$

В самом деле, предположим, для простоты, что X принимает только два значения x_1 и x_2 с вероятностями p_1 и p_2 , а Y — только два значения y_1 и y_2 с вероятностями g_1 и g_2 . Обозначим A_i событие $X = x_i$, а B_k — событие $Y = y_k$. Пусть $p_{ik} = P(A_i B_k)$ — вероятность совмещения $A_i B_k$. Тогда для $M(X + Y)$ можем написать

$$\begin{aligned} M(X + Y) &= (x_1 + y_1)p_{11} + (x_1 + y_2)p_{12} + (x_2 + y_1)p_{21} + \\ &+ (x_2 + y_2)p_{22} = x_1(p_{11} + p_{12}) + x_2(p_{21} + p_{22}) + \\ &+ y_1(p_{11} + p_{21}) + y_2(p_{12} + p_{22}). \end{aligned} \quad (2)$$

Так как события B_1 и B_2 несовместны и составляют полную группу, события $A_1 B_1$ и $A_1 B_2$ также несовместны. Поэтому по формуле сложения

$$\begin{aligned} p_{11} + p_{12} &= P(A_1 B_1) + P(A_1 B_2) = P(A_1 B_1 + A_1 B_2) = \\ &= P[A_1(B_1 + B_2)] = P(A_1 U) = P(A_1) = p_1. \end{aligned}$$

Аналогично доказывается, что $p_{21} + p_{22} = p_2$, $p_{11} + p_{21} = g_1$ и $p_{12} + p_{22} = g_2$. Подставив эти значения в (2), получим

$$M(X + Y) = x_1 p_1 + x_2 p_2 + y_1 g_1 + y_2 g_2 = M(X) + M(Y),$$

что и требовалось.

Свойство 3° распространяется на любое число слагаемых, т. е. $M\left(\sum_{i=1}^m X_i\right) = \sum_{i=1}^m M(X_i)$.

4°. Математическое ожидание произведения независимых случайных величин равно произведению их математических ожиданий:

$$M(XY) = M(X) \cdot M(Y).$$

Доказательство проводится так же, как для суммы, но упрощается из-за независимости X и Y . В этом случае, как мы знаем, $p_{ik} = p_i g_k$.

Свойство 4° распространяется на любое число сомножителей, т. е. $M(X_1 X_2 \dots X_n) = M(X_1) M(X_2) \dots M(X_n)$.

Из свойств 3° и 4° следует, что математическое ожидание линейно:

$$M\left(\sum_{i=1}^n C_i X_i\right) = \sum_{i=1}^n C_i M(X_i).$$

Пример. Пусть X — число выпавших гербов при бросании первой монеты, а Y — второй. Тогда

$$M(X) = M(Y) = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2},$$

$$M(X + Y) = M(X) + M(Y) = \frac{1}{2} + \frac{1}{2} = 1,$$

$$M(XY) = M(X) \cdot M(Y) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Математическое ожидание называют иногда *центром распределения* значений случайной величины. Этот термин появился по аналогии с задачей механики, в которой координата центра масс m_1, m_2, \dots, m_n , имеющих координаты x_1, x_2, \dots, x_n , вычисляется по формуле

$$\bar{x} = \frac{\sum_{i=1}^n x_i m_i}{\sum_{i=1}^n m_i}.$$

Обозначим $m = \sum_{i=1}^n m_i$, $p_i = \frac{m_i}{m}$. Очевидно, что $\sum_{i=1}^n p_i = 1$.

Формула для \bar{x} принимает вид

$$\bar{x} = \sum_{i=1}^n x_i p_i,$$

что вполне аналогично формуле для $M(X)$.

3. Дисперсия

Математическое ожидание случайной величины дает важную информацию о центре, вокруг которого группируются значения этой величины, но не отвечает на вопрос о том, как плотно эти значения примыкают к центру. Так, две случайные величины X и Y , законы распределения которых представлены табл. 20 и 21, имеют одно и то же математическое ожидание. Однако в первом случае разброс от нуля в 100 раз больше, чем во втором. Чтобы учесть этот разброс, или, как говорят, рассеяние, вводят специальные числовые характеристики случайной величины.

Таблица 20

X	-100	0	100
P	1/4	1/2	1/4

Таблица 21

X	-1	0	1
P	1/4	1/2	1/4

Пусть случайная величина X принимает значения x_i с вероятностями p_i . Рассмотрим разность $X - M(X)$. Эту разность называют *отклонением* случайной величины X от ее центра. Очевидно, что $X - M(X)$ является случайной величиной со значениями $x_i - M(X)$ и вероятностями p_i . На первый взгляд может показаться, что рассеяние X можно охарактеризовать средним значением отклонения, т. е. математическим ожиданием величины $X - M(X)$. Однако это не так. Легко проверить, что математическое ожидание отклонения равно нулю. В самом деле, в силу свойств 1°—3° имеем $M(X - M(X)) = M(X) - M(M(X)) = M(X) - M(X) = 0$. Это и понятно. Математическое ожидание потому и названо центром, что отклонения в одну сторону уравновешиваются отклонениями в другую. Указанное обстоятельство вынуждает искать другие характеристики рассеяния. Одной из них является дисперсия.

Определение. Дисперсией дискретной случайной величины называется математическое ожидание квадрата ее отклонения. Обозначается дисперсия $D(X)$.

Таким образом,

$$D(X) = M(X - M(X))^2. \quad (3)$$

Поскольку случайная величина $(X - M(X))^2$ принимает свои значения $(x_i - M(X))^2$ с вероятностями p_i , равенство (3) подробно можно записать в виде

$$D(X) = (x_1 - M(X))^2 p_1 + (x_2 - M(X))^2 p_2 + \dots \\ \dots + (x_n - M(X))^2 p_n. \quad (4)$$

Если $(x_i - M(X))^2 = (x_k - M(X))^2$, то их общему значению, как обычно, приписывается вероятность $p_i + p_k$.

Из равенства (4) видно, что дисперсия $D(X)$ действительно может служить характеристикой рассеяния. Если $D(X)$ мала, то малыми должны быть все слагаемые справа в (4). Это означает, что если какое-нибудь значение отклонения $x_k - M(X)$ велико, то соответствующая вероятность p_k должна быть мала.

Для примера найдем дисперсию случайных величин X и Y , приведенных в начале раздела. Имеем

$$D(X) = (\pm 100 - 0)^2 \left(\frac{1}{4} + \frac{1}{4} \right) + (0 - 0)^2 \cdot \frac{1}{2} = \\ = \frac{1}{2} \cdot 10000 = 5000,$$

$$D(Y) = (\pm 1 - 0)^2 \left(\frac{1}{4} + \frac{1}{4} \right) + (0 - 0)^2 \cdot \frac{1}{2} = 0,5.$$

Пользуясь определением дисперсии и свойствами математического ожидания, установим свойства дисперсии.

1°. Справедливая формула

$$D(X) = M(X^2) - [M(X)]^2. \quad (5)$$

Таким образом, дисперсия случайной величины X равна математическому ожиданию квадрата X без квадрата ее математического ожидания.

В самом деле, пользуясь формулой сокращенного умножения для случайных величин, можем написать

$$D(X) = M[X - M(X)]^2 = M[X^2 - 2XM(X) + M^2(X)] = \\ = M(X^2) - 2M(X)M(X) + M^2(X) = M(X^2) - M^2(X),$$

что и требовалось.

Формула (5) намного удобнее для вычисления дисперсии, чем равенство (4). Ею, как правило, и пользуются на практике.

2°. Дисперсия постоянной равна нулю:

$$D(C) = 0.$$

В самом деле, пользуясь формулой (5), можно написать $D(C) = M(C^2) - [M(C)]^2 = C^2 - C^2 = 0$. Смысл этого утверждения очевиден: постоянная не имеет разброса.

3°. Постоянный множитель можно выносить за знак дисперсии, возводя его в квадрат:

$$D(CX) = C^2 D(X).$$

Действительно, по формуле (5) $D(CX) = M(C^2 X^2) - [M(CX)]^2 = C^2 M(X^2) - C^2 [M(X)]^2 = C^2 D(X)$. Таким образом, умножение случайной величины на постоянную увеличивает рассеяние (при $|C| > 1$) или уменьшает его (при $|C| < 1$). Это обстоятельство учитывается при изменении масштаба копструкций.

4°. Дисперсия суммы или разности двух независимых случайных величин равна сумме их дисперсий:

$$D(X \pm Y) = D(X) + D(Y).$$

В самом деле:

$$\begin{aligned} D(X - Y) &= M(X - Y)^2 - [M(X - Y)]^2 = \\ &= M(X^2 - 2XY + Y^2) - (M(X) - M(Y))^2 = \\ &= M(X^2) - 2M(X)M(Y) + M(Y^2) - M^2(X) + \\ &\quad + 2M(X)M(Y) - M^2(Y) = D(X) + D(Y). \end{aligned}$$

Свойство 4° распространяется на любое число слагаемых. Следствием свойств 2° и 4° является.

5°. Дисперсия суммы случайной величины X и постоянной C равна дисперсии X :

$$D(X + C) = D(X).$$

Во многих практических и теоретических вопросах вместо дисперсии, которая имеет размерность квадрата случайной величины, удобнее рассматривать корень квадратный из дисперсии. Эту характеристику называют *средним квадратичным отклонением* и обозначают $\sigma(X)$. Таким образом,

$$\sigma(X) = \sqrt{D(X)}.$$

Очевидно, $\sigma(X)$ имеет размерность случайной величины. Поэтому если случайная величина измеряется, например, в граммах, то и рассеяние ее, характеризующее $\sigma(X)$, будет так же измеряться в граммах, а не в квадратных граммах.

Для случайных величин X и Y , рассмотренных в начале раздела, имеем $\sigma(X) = \sqrt{D(X)} = 100/\sqrt{2}$ и $\sigma(Y) = 1/\sqrt{2}$. Таким образом, среднее квадратичное отклонение X в 100 раз больше, чем у Y .

Наряду с распространенными числовыми характеристиками $M(X)$ и $D(X) = M(X - M(X))^2$ употребляются и другие, например $M(X^k)$ и $M(X - M(X))^k$. Первые из них называются начальными, а вторые — центральными моментами порядка k . В этих терминах $M(X)$ — начальный момент первого, а $D(X)$ — центральный момент второго порядка.

4. Числовые характеристики биномиального распределения

Со схемой Бернулли, как мы знаем, связаны случайные величины двух типов: X_i со значениями 1, 0 и вероятностями p , $q = 1 - p$ и $X = \sum_{i=1}^n X_i$ со значениями 0, 1, 2, ..., m , ..., n и вероятностями $P_n(m)$. Найдем числовые характеристики этих случайных величин.

Для X_i имеем

$$M(X_i) = 1 \cdot p + 0 \cdot q = p,$$

$$\begin{aligned} D(X_i) &= M(X_i^2) - [M(X_i)]^2 = 1 \cdot p + 0 \cdot q - p^2 = \\ &= p - p^2 = p(1 - p) = p \cdot q, \end{aligned}$$

а для X —

$$M(X) = M\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n M(X_i) = p \cdot n,$$

$$D(X) = D\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n D(X_i) = p \cdot q \cdot n.$$

Итак, $M(X) = p$, $D(X) = p \cdot q$, $M(X) = p \cdot n$, $D(X) = p \cdot q \cdot n$, $\sigma(X) = \sqrt{p \cdot q \cdot n}$,

Мы неоднократно будем пользоваться этими выражениями в дальнейшем.

Пример. Вероятность изготовления бракованной детали $p = 0,015$. Сколько бракованных деталей можно ожидать в партии из 1200 штук?

Мы уже отвечали на подобные вопросы, опираясь на интуитивные представления о вероятности как о доли интересующих нас событий. Сейчас мы имеем возможность решать подобные задачи строго, не прибегая к интуиции. Будем рассматривать проверку деталей на наличие брака как серию испытаний схемы Бернулли, в которой число испытаний $n = 1200$, а вероятность наступления события A в одном испытании, т. е. вероятность обнаружения брака, $p = 0,015$. Случайная величина X в этой схеме — это число бракованных деталей в партии. Тогда среднее число бракованных деталей — это $MX = np = 1200 \cdot 0,015 = 18$. Таким образом, число бракованных деталей, меняясь от партии к партии, в среднем равно 18. Рассеяние от этого среднего значения характеризуется дисперсией $D(X) = n \cdot p \cdot q = 1000 \cdot 0,015 \cdot 0,985 = 14,775$ и средним квадратичным отклонением $\sigma(X) = \sqrt{D(X)} = \sqrt{14,775} \approx 3,85$.

5. Числовые характеристики распределения Пуассона

До сих пор мы рассматривали дискретные случайные величины с конечным множеством значений. В случае, когда случайная величина X принимает счетное множество значений $x_1, x_2, \dots, x_n, \dots$ с вероятностями $p_1, p_2, \dots, p_n, \dots$, ее математическое ожидание определяется формулой

$$M(X) = \sum_{k=1}^{\infty} x_k p_k. \quad (6)$$

Здесь справа написана сумма сходящегося ряда.

Определение дисперсии остается неизменным: $D(X) = M(X - MX)^2$. Сохраняются и свойства этих числовых характеристик.

Примером дискретной случайной величины со счетным множеством значений является величина X , распределенная по закону Пуассона. Значения 0, 1, 2, ...

..., k , ... она принимает с вероятностями $P_k = P(k) = \frac{a^k}{k!} e^{-a}$. Поэтому по формуле (6) можем написать

$$\begin{aligned} M(X) &= \sum_{k=0}^{\infty} k P(k) = \sum_{k=0}^{\infty} k \frac{a^k e^{-a}}{k!} = \sum_{k=1}^{\infty} \frac{a^k e^{-a}}{k!} = \\ &= a e^{-a} \sum_{k=1}^{\infty} \frac{a^{k-1}}{(k-1)!} = a e^{-a} \sum_{s=0}^{\infty} \frac{a^s}{s!} = a \cdot e^{-a} \cdot e^a = a. \end{aligned}$$

Мы, наконец, выяснили смысл параметра a в формуле пуассоновского распределения. Это математическое ожидание, т. е. среднее значение случайной величины, распределенной по пуассоновскому закону. Теперь стало понятно, почему, заменяя биномиальное распределение пуассоновским, мы полагали $np = a$. Пуассоновское приближение при этом имело то же среднее значение, что и исходная биномиальная случайная величина.

Найдем теперь дисперсию $D(X)$. Имеем

$$\begin{aligned} M(X^2) &= \sum_{k=0}^{\infty} k^2 P(k) = \sum_{k=0}^{\infty} [k(k-1) + k] P(k) = \\ &= \sum_{k=2}^{\infty} k(k-1) \frac{a^k e^{-a}}{k!} + \sum_{k=0}^{\infty} k \frac{a^k e^{-a}}{k!} = a^2 e^{-a} \sum_{k=2}^{\infty} \frac{a^{k-2}}{(k-2)!} + \\ &+ a = a^2 \cdot e^{-a} \cdot e^a + a = a^2 + a. \end{aligned}$$

Поэтому

$$D(X) = M(X^2) - [M(X)]^2 = a^2 + a - a^2 = a.$$

Таким образом, дисперсия случайной величины, распределенной по закону Пуассона, совпадает с математическим ожиданием этой величины и равна параметру в формуле закона Пуассона.

Этой замечательной особенностью пуассоновского распределения пользуются на практике. Если какое-то распределение «подозревается» в пуассоновости, то прежде всего проверяется выполнение равенства $M(X) = D(X)$. В случае его выполнения (хотя бы приближенно) составляется распределение Пуассона $P(k) = \frac{a^k e^{-a}}{k!}$, где $a = M(X) = D(X)$, и это распределение сравнивается с исследуемым.

Т а б л и ц а 22

X	0	1	2	3	4
P	0,52	0,33	0,125	0,02	0,005

Пример. Родственные предприятия отрасли в течение 20 лет обследуются по поводу тяжелого травматизма. Всего обследовано 50 предприятий и, таким образом, проведено $n = 20 \cdot 50 = 1000$ наблюдений. В каждом из них зафиксировано какое-то число тяжелых травм на определенном предприятии в определенном году. При этом выяснилось, что в 520 наблюдениях не отмечено ни одного случая тяжелого травматизма, в 330 отмечен 1 случай, в 125 наблюдениях по 2 случая, в 20 по 3 и в 5 наблюдениях по 4 случая тяжелого травматизма.

Пусть X — случайная величина, значения которой равны числу несчастных случаев в одном наблюдении. Из сказанного следует, что этими значениями являются 0, 1, 2, 3 и 4. Вероятности, с которыми X принимает свои значения, строго говоря, нам неизвестны. Однако, поскольку наблюдений довольно много, вместо вероятностей можно взять относительные частоты. Тогда можно написать закон распределения X (табл. 22). Пользуясь этим законом, найдем $M(X)$ и $D(X)$. Имеем

$$M(X) = 0 \cdot 0,52 + 1 \cdot 0,33 + 2 \cdot 0,125 + \\ + 3 \cdot 0,02 + 4 \cdot 0,005 = 0,66,$$

$$M(X^2) = 1 \cdot 0,33 + 4 \cdot 0,125 + 9 \cdot 0,02 + 16 \cdot 0,005 = 1,09.$$

Поэтому $D(X) = M(X^2) - (MX)^2 \approx 1,09 - 0,44 = 0,65$.

Видно, что $D(X)$ мало отличается от MX . Это дает основание предположить, что случайная величина распределена по закону Пуассона $P(k) = \frac{a^k}{k!} e^{-a}$ с параметром $a = 0,66$. Отыскав в таблице значения $p_a(k)$ при $a = 0,66$ и $k = 0, 1, 2, 3, 4$, обнаружим, что $P(0) = 0,5$, $P(1) = 0,34$, $P(2) = 0,120$, $P(3) = 0,02$, $P(4) = 0,005$. Это вполне удовлетворительно совпадает со второй строчкой табл. 22 и говорит о том, что построенное распре-

деление Пуассона может быть использовано для характеристики числа тяжелых травм на всех предприятиях отрасли, подобных тем, которые были обследованы. В частности, можно получить ответ на вопрос, какова вероятность 5 несчастных случаев в год. По той же таблице $p_a(k)$ находим $P(5) = 0,0007$. Как видим, эта вероятность практически равна нулю. Таким образом, если на каком-то предприятии произошло 5 или более несчастных случаев в году, то это сигнал о серьезных нарушениях техники безопасности. Аналогично решается вопрос, если на предприятии два года подряд наблюдалось по 3 или 4 несчастных случая. В распределении Пуассона этим событиям соответствуют хотя и не нулевые, но довольно малые вероятности. Вероятность же их совмещения, равная произведению вероятностей, исчезающе мала.

6. Математическое ожидание в генетике

Мы уже указывали на то, какую важную роль играет математическое ожидание во всех практических задачах, связанных со случайными величинами. Рассмотрим теперь несколько задач из области генетики, в решении которых существенно используется математическое ожидание распределения Бернулли.

Генетический дрейф. Пусть в родительской популяции, содержащей три генотипа AA , Aa и aa , частоты генов A и a равны соответственно p и q . Рассмотрим популяцию потомства. Пусть популяция потомства состоит из N особей. Перепούμε эти особи и выстроим их по их порядковым номерам. Получим последовательность вида $AAaAaaaaAaAAaA$, которую мы можем рассматривать как слово длиной $2N$, образованное из алфавита $\{A, a\}$. Будем испытывать позиции этого слова по схеме Бернулли, считая, что испытание окончилось удачей, если на испытываемой позиции стоит буква A , и неудачей, если буква a . Случайная величина X этой схемы — число букв A в слове. Вообще говоря, X может принимать значения $0, 1, 2, \dots, 2N$. Их вероятности вычисляются по формуле Бернулли: $P(k) = C_{2N}^k p^k q^{2N-k}$.

Вместо X рассмотрим случайную величину $\frac{X}{2N}$. Очевидно, $\frac{X}{2N}$ — это относительная частота буквы A в сло-

ве, т. е. относительная частота гена A в популяции. Таким образом, относительная частота гена A при переходе от родительской популяции к потомству не остается неизменной. С вероятностью $P(k)$ она может принять любое значение:

$$\frac{0}{2N}, \frac{1}{2N}, \dots, \frac{k}{2N}, \dots, \frac{2N}{2N}.$$

Этот вывод как будто противоречит закону Харди — Вайнберга. Но на самом деле противоречия нет. Результат, который мы получили, относится к одной фиксированной популяции. Если же популяций много или одна большая популяция разбита на части, то надо говорить не о величине $\frac{X}{2N}$, а о ее среднем значении, т. е. о математическом ожидании. Для биномиального распределения имеем $M(X) = np = 2Np$. Поэтому в силу свойства математического ожидания $M\left(\frac{X}{2N}\right) = \frac{1}{2N} M(X) = \frac{1}{2N} 2Np = p$. Это вполне согласуется с законом Харди — Вайнберга.

Таким образом, при переходе от поколения к поколению относительная частота гена A флуктуирует, т. е. случайно колеблется, принимая с вероятностью $P(k)$ значения $\frac{k}{2N}$, $k = 0, 1, 2, \dots, 2N$. Однако в среднем эта частота равна p . Аналогичный вывод справедлив, очевидно, и для гена a . Это явление называется *дрейфом генов*.

Отклонения от среднего значения описываются дисперсией

$$\sigma^2 = D\left(\frac{X}{2N}\right) = \frac{1}{4N^2} D(X) = \frac{1}{4N^2} \cdot 2Npq = \frac{pq}{2N}.$$

Как видно, чем больше численность популяции N , тем меньше дисперсия.

Можно попробовать найти вероятность сохранения заданной относительной частоты p в новом поколении. Например, при $p = \frac{1}{2}$ имеем $q = \frac{1}{2}$. Когда $\frac{X}{2N}$ принимает значение p , т. е. когда $\frac{X}{2N} = \frac{1}{2}$, имеем $X = N$. По

формуле Бернулли при $k = N$ можем написать

$$P(N) = C_{2N}^N p^N q^{2N-N} = \frac{1}{2^{2N}} C_{2N}^N.$$

Эта формула уже встречалась нам при решении задачи о шахматном матче примерно равных по силе шахматистов. При большом N вычисления очень трудны. Приближением Пуассона воспользоваться нельзя, так как p заметно отличается от нуля. В следующих главах, где рассматриваются непрерывные случайные величины, мы укажем способ приближенной оценки $P(k)$ в подобных случаях.

«Сборка» молекулы белка. В разделе, посвященном приложениям комбинаторики в генетике, мы установили необычайно большое количество различных слов-молекул, которые можно образовать из алфавита нуклеотидов или аминокислот. Если бы сборка таких молекул происходила случайно, то вероятность образования заданной молекулы, например белка (слова длиной 100 из алфавита мощностью 20), было бы равна $1:20^{100} \approx 1:10^{130}$. Другими словами, если бы реакция сборки одной такой молекулы длилась одну секунду, то в среднем пришлось бы ждать 10^{130} секунд, пока случайно ли образуется нужная молекула. Это во много-много раз больше 10^{17} секунд — времени, которым оценивается возраст Земли.

К счастью, молекулы белков и нуклеиновых кислот собираются не вполне случайно. Точнее говоря, сборка происходит на молекуле-матрице. Если по-прежнему говорить о сборке молекулы белка, то можно представить такую схему. В первый момент все 100 позиций матрицы свободны. К каждой позиции случайно подходит одна из 20 букв. Однако сила химической связи между различными буквами и позициями неодинакова. Каждая позиция может прочно удерживать только одну определенную букву. Такую букву по отношению к данной позиции будем называть правильной. Правильная буква, подойдя к своей позиции, фиксируется на ней. Неправильная буква быстро покидает позицию, и позиция снова становится свободной. Предположим для простоты, что все буквы подходят к позициям одновременно и узнавание правильных или неправильных букв происходит в течение одной секунды. По прошествии этой секунды часть

позиций будет занята правильными буквами, а другая часть останется свободной. Так как вероятность, с которой к данной позиции подходит правильная буква, равна $p = 1/20$ и не зависит от того, какие буквы подошли к другим позициям, случайная величина X — число занятых позиций — распределена по биномиальному закону. В таком случае среднее число занятых позиций, т. е. математическое ожидание MX , вычисляется по формуле $MX = np = 100 \cdot \frac{1}{20} = 5$.

Отсюда следует, что к началу второй секунды незанятыми (в среднем) останутся лишь $n_1 = n - n \cdot p = = n \cdot q$ позиций ($q = 1 - p$). К каждой из них снова случайно подойдут буквы и займут в среднем $n_1 p = = n \cdot p \cdot q$ позиций, оставляя свободными $n_2 = n_1 - - n_1 p = n_1 \cdot q = n \cdot q^2$ позиций. Подобным образом в течение третьей секунды будет занято еще $n_2 \cdot p = = n \cdot p \cdot q^2$ позиций, в течение четвертой секунды — еще $n \cdot p \cdot q^3$ позиций, наконец, после m -й секунды прибавится еще $n \cdot p \cdot q^{m-1}$ занятых позиций. Всего после m секунд будет занято

$$n \cdot p + n \cdot p \cdot q + n \cdot p \cdot q^2 + \dots + n \cdot p \cdot q^{m-1}$$

позиций. Это сумма S_m первых m членов геометрической прогрессии с первым членом $n \cdot p$ и знаменателем q . Как известно,

$$S_m = \frac{np(1 - q^m)}{1 - q} = n(1 - q^m).$$

Найдем такое m , чтобы после m секунд в среднем осталось меньше одной свободной позиции, т. е. чтобы $n - S_m < 1$. Имеем $n - n(1 - q^m) < 1$. Отсюда $n \cdot q^m < 1$ и, следовательно, $m > -\frac{\lg n}{\lg q}$. В нашем случае $\lg n = = \lg 100 = 2$, $\lg q = \lg(1 - p) = \lg \frac{19}{20} = -0,0222$. Поэтому $m > \frac{2}{0,0222} \approx 90$.

Таким образом, на сборку заданной белковой молекулы на матрице требуется не более полутора минут. Это, конечно, не идет ни в какое сравнение с теми астрономическими числами, которые получились без учета матричных связей.

Среднее число повторов в генах. Одна из характерных особенностей первичной структуры

генов — наличие в ней повторов. В длинной последовательности нуклеотидов, т. е. в слове длиной N , образованном из алфавита $\{T, A, C, G\}$, можно обнаружить пары повторяющихся фрагментов длиной l . Иногда фрагменты в паре повторяются во всех позициях (с точностью до комплементарности). Часто, однако, некоторые позиции в достаточно длинных фрагментах различаются.

Случайны ли такие повторы? Встречаются ли они просто потому, что мощность алфавита невелика, а длина слова N большая? Или их наличие в структуре гена играет какую-то специальную роль, и поэтому число повторов в реальных нуклеотидных последовательностях заметно отличается от среднего числа повторов, которые могут образоваться случайно? Чтобы ответить на эти вопросы, нужно знать среднее число случайных повторов.

Рассмотрим случайно образованную последовательность длиной N . Выделим в ней какой-нибудь фрагмент длиной l . Назовем его первым. Правее этого фрагмента может быть расположен совпадающий с ним (полностью или частично) другой фрагмент. Назовем его вторым. Какова вероятность того, что второй фрагмент в k позициях отличается от первого?

На i -й позиции второго фрагмента, вообще говоря, может оказаться любая буква. С некоторой вероятностью p эта буква совпадает с той, которая стоит на i -й позиции в первом фрагменте. В простейшем случае, если не различать комплементарных букв, можем считать $p = 1/2$. В более сложных случаях, учитывающих силу химических связей, тепловые помехи и т. п., p может отличаться от $1/2$, но мы всегда будем считать эту вероятность одной и той же для всех позиций. Тогда вероятность того, что второй фрагмент ровно в k позициях совпадает с первым, может быть найдена по формуле Бернулли

$$P = P_l(k) = C_l^h \cdot p^h \cdot q^{l-h}, \quad q = 1 - p.$$

Мы нашли вероятность образования повтора, т. е. пары повторяющихся друг друга фрагментов (с $l - k$ различиями) при некотором фиксированном расположении фрагментов. Сколько же всего таких возможных расположений?

Первый фрагмент может быть расположен $N - 2l + 1$ способами. В самом деле, будем следить за первой буквой первого фрагмента. Так как после этой буквы должны поместиться еще $l - 1$ букв первого фрагмента, а затем l букв второго фрагмента (он всегда правее первого), для первой буквы первого фрагмента запрещены $l + l - 1$ мест на правом краю исходной последовательности. Остальные места для этой буквы доступны, и всего таких мест $N - (l + l - 1) = N - 2l + 1$. Это и есть число способов, которыми может быть размещен первый фрагмент.

Пусть первая буква первого фрагмента находится на i -й позиции исходной последовательности, $1 \leq i \leq N - 2l + 1$. Найдем число способов размещения второго фрагмента при этом i . По условию для первой буквы второго фрагмента запрещены $i - 1$ мест слева от начала первого фрагмента. Кроме того, ей недоступны l мест, занятых вторым фрагментом. Наконец, она не может занять ни одно из $l - 1$ мест на правом краю исходной последовательности (иначе на этом краю не поместится второй фрагмент). Остальные места для первой буквы второго фрагмента доступны, и всего таких мест $N - (i - 1) - l - (l - 1) = N - i - 2l + 2$. Это и есть число способов размещения второго фрагмента при заданном i , т. е. при заданном положении первого фрагмента.

Суммируя по всем возможным i от 1 до $N - 2l + 1$, получим число n всех возможных расположений двух непересекающихся фрагментов длиной l :

$$n = (N - 1 - 2l + 2) + (N - 2 - 2l + 2) + \dots + [N - (N - 2l + 1) - 2l + 2].$$

Это сумма $N - 2l + 1$ первых членов арифметической прогрессии с первым членом $N - 2l + 1$ и разностью $d = -1$. Последний член в ней равен $N - (N - 2l + 1) - 2l + 2 = 1$. Поэтому по известной формуле

$$\begin{aligned} n &= \frac{(N - 2l + 1) + 1}{2} (N - 2l + 1) = \\ &= \frac{(N - 2l + 1)(N - 2l + 2)}{2}. \end{aligned}$$

Итак, имеется n всевозможных расположений фрагментов. При каждом таком расположении с одной и той же вероятностью P может образоваться повтор.

Следовательно, случайная величина X — число повторов во всей серии n расположений — распределена по биномиальному закону. Ее среднее значение MX , т. е. среднее число повторов, как известно, равно произведению nP . Таким образом, среднее число повторов фрагментов длиной l с $l-k$ различиями в произвольной последовательности длиной N вычисляется по формуле

$$MX = np = \frac{(N - 2l + 1)(N - 2l + 2)}{2} C_e^k \cdot p^k \cdot q^{l-k}.$$

Видно, что MX растет с ростом N .

Сравнив это число с числом повторов в реальных нуклеотидных последовательностях, можно сделать вывод о том, насколько случайны подобные повторы. Если число повторов в реальных последовательностях примерно равно MX , то, по-видимому, повторы носят случайный, непредопределенный характер. Если же число реальных повторов заметно отличается от MX , то можно предположить, что в основе их формирования лежит неслучайный механизм.

Обычно интересны повторы с небольшим числом различий. Если, например, различаются только две позиции, то $k = l - 2$, и при $p = q = 1/2$ формула для MX принимает простой вид:

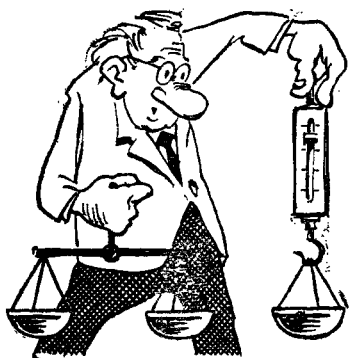
$$MX = \frac{(N - 2l + 1)(N - 2l + 2)l(l - 1)}{2^{l+2}}.$$

Интересно исследовать величину MX в зависимости от длины фрагмента l .

Задача, которую мы сейчас рассмотрели, относится к новому разделу генетики — анализу генетических текстов. Интенсивные исследования в этой области ведутся в Институте цитологии и генетики Сибирского отделения АН СССР. О них можно прочесть, например, в книге В. А. Ратнера и др. «Проблемы теории молекулярной эволюции» (Новосибирск: Наука, 1985).

Непрерывная случайная величина

1. Закон распределения непрерывной случайной величины



Как мы уже условились, непрерывной случайной величиной мы называем такую случайную величину, которая может принимать любые значения из некоторого промежутка. При изучении таких величин мы можем опереться на известные факты из анализа функций действительных переменных, и это часто быстрее приводит к успеху, чем в случае дискретных случайных величин.

Функция распределения непрерывной случайной величины. Так же, как и в случае дискретных случайных величин, мы должны научиться задавать закон распределения вероятностей непрерывной случайной величины. Если мы попытаемся копировать те приемы, которые использовались в дискретном случае, то вряд ли добьемся успеха. Ведь значения непрерывной случайной величины сплошь заполняют некоторый промежуток или даже всю числовую ось. Каждому значению мы должны были бы приписать вероятность и проследить за тем, чтобы сумма этих вероятностей была равна единице. Но как образовать подобную сумму?

С другой стороны, геометрические соображения приводят к выводу, что вероятность любого отдельно взятого значения равна нулю. В самом деле, вычисляя вероятность по геометрической схеме, мы должны длину промежутка благоприятствующих точек разделить на длину промежутка всех возможных значений. А длина единственной точки, соответствующей отдельному значению случайной величины, равна нулю.

Указанные трудности заставляют нас отказаться от попытки следить за вероятностями отдельных значений непрерывной случайной величины. Вместо этого будем задавать вероятность целых промежутков значений. Точнее говоря, будем задавать вероятность события, состоящего в том, что значения случайной величины лежат в данном промежутке.

Аналогичные соображения применяются при построении моделей в механике. Масса отдельной точки равна нулю. Но масса стержня, заполненного точками, отлична от нуля.

Итак, мы хотим описать непрерывную случайную величину с помощью вероятностей промежутков ее значений.

Рассмотрим прежде всего промежуток $(-\infty; x)$. Обозначим $P(X < x)$ вероятность того, что значения случайной величины X лежат в этом промежутке. Очевидно, что вероятность $P(X < x)$, вообще говоря, зависит от x , т. е. $P(X < x)$ является функцией от x . Эта функция обозначается $F(x)$ и называется *функцией распределения* непрерывной случайной величины.

Из общих свойств вероятности следуют некоторые свойства функции распределения $F(x) = P(X < x)$.

1°. Значения функции $F(x)$ лежат между 0 и 1:

$$0 \leq F(x) \leq 1.$$

В самом деле, по определению $F(x)$ — вероятность, а значения вероятности лежат между 0 и 1.

2°. Функция $F(x)$ — неубывающая:

$$F(x_2) \geq F(x_1), \text{ если } x_2 > x_1.$$

Действительно, пользуясь формулой сложения вероятностей

$$P(X < x_2) = P(X < x_1) + P(x_1 \leq X < x_2),$$

можем написать

$$\begin{aligned} F(x_2) - F(x_1) &= P(X < x_2) - P(X < x_1) = \\ &= P(x_1 \leq X < x_2) \geq 0, \end{aligned} \quad (1)$$

так как любая вероятность неотрицательна.

3°. Имеет место равенство

$$\lim_{x \rightarrow \infty} F(x) = F(\infty) = 1.$$

В самом деле, будучи монотонной и ограниченной, $F(x)$ имеет предел $F(\infty) = \lim_{x \rightarrow \infty} F(x)$. По определению $F(\infty) = P(x < \infty) = 1$, так как это вероятность достоверного события.

4°. Имеет место равенство

$$\lim_{x \rightarrow -\infty} F(x) = F(-\infty) = 0.$$

Действительно, монотонность и ограниченность обеспечивают существование предела $F(-\infty) = \lim_{x \rightarrow -\infty} F(x)$. С другой стороны, по определению $F(-\infty) = P(X < -\infty) = 0$, так как это вероятность невозможного события.

5°. Имеет место равенство

$$P(a \leq X < b) = F(b) - F(a). \quad (2)$$

Это равенство следует из формулы (1), если в ней положить $x_1 = a$ и $x_2 = b$.

Поскольку вероятность одного значения равна нулю, вместе с равенством (2) можем написать

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X \leq b) = P(a \leq X < b) = \\ &= P(a < X < b) = F(b) - F(a). \end{aligned} \quad (3)$$

Таким образом, вероятность того, что непрерывная случайная величина X принимает значения из некоторого промежутка, равна приращению на этом промежутке ее функции распределения.

Вообще говоря, можно указать случайные величины, функция распределения которых $F(x) = P(X < x)$ лишь кусочно-непрерывна. Но мы ограничимся практически наиболее важным случаем, когда $F(x)$ непрерывна. Из свойств 1°—5° следует, что в этом случае график $F(x)$ — непрерывная неубывающая кривая, заключенная между горизонтальными асимптотами $y = 0$ и $y = 1$ (рис. 13).

Плотность распределения непрерывной случайной величины. Во многих теоретических построениях и практических задачах функция

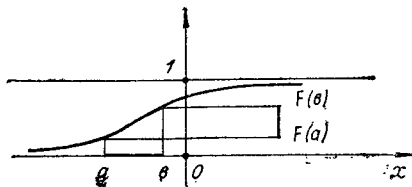


Рис. 13.

распределения $F(x)$ не только непрерывна, но даже дифференцируема или по крайней мере кусочно дифференцируема. В таких случаях непрерывную случайную величину X удобно описывать с помощью производной функции $F(x)$. Рассмотрим этот вопрос.

Пусть случайная величина X имеет дифференцируемую функцию распределения $F(x)$. Производная этой функции обозначается $f(x)$ и называется *плотностью распределения* вероятностей случайной величины X . Таким образом,

$$f(x) = F'(x). \quad (4)$$

Равенство (4), очевидно, эквивалентно равенству

$$F(x) = \int_{-\infty}^x f(x) dx \quad (5)$$

и означает, что $F(x)$ — первообразная плотности $f(x)$.

Термин «плотность» объясняется аналогией с механикой. Если $M(x)$ — масса участка стержня от его левого конца до точки x , то плотность стержня $\rho(x)$ с массой $M(x)$ связана, как известно, равенством

$$\rho(x) = M'(x) = \lim_{\Delta x \rightarrow 0} \frac{M(x + \Delta x) - M(x)}{\Delta x},$$

где $M(x + \Delta x) - M(x)$ есть масса участка стержня от x до $x + \Delta x$. Аналогично этому в нашем случае

$$f(x) = F'(x) = \lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x}.$$

Переходя к вероятности и пользуясь формулой (3), последнее равенство можно написать в виде

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}.$$

Это равенство объясняет вероятностный смысл плотности $f(x)$.

Функцию $f(x)$ называют также *законом распределения* вероятностей случайной величины X . Из свойств 1°—5° функции распределения $F(x)$ вытекают свойства плотности $f(x)$.

1°. *Плотность неотрицательна:*

$$f(x) \geq 0.$$

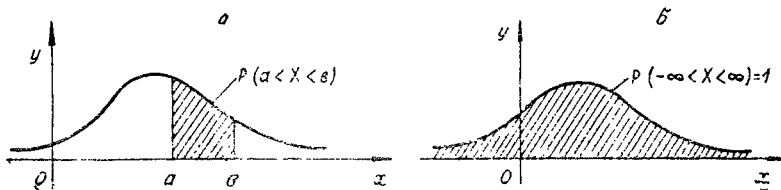


Рис. 14.

В самом деле, функция $F(x)$ не убывает. Поэтому ее производная $f(x)$ не может быть отрицательной.

2°. Интеграл от плотности по всей оси равен 1:

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

Действительно,

$$\int_{-\infty}^{\infty} f(x) dx = F(\infty) - F(-\infty) = 1.$$

3°. Имеет место равенство

$$\int_{-\infty}^{\infty} f(x) dx = P(X < x). \quad (6)$$

В самом деле,

$$\int_{-\infty}^x f(x) dx = F(x) - F(-\infty) = P(X < x).$$

Равенство (6), по сути совпадающее с (5), дает интегральное представление вероятности $P(X < x)$ (рис. 14, а)

4°. Имеет место равенство

$$\int_a^b f(x) dx = P(a \leq x \leq b). \quad (7)$$

Действительно,

$$\int_a^b f(x) dx = F(b) - F(a) = P(a \leq x \leq b)$$

в силу (3).

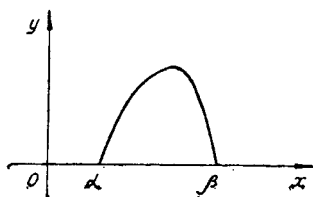


Рис. 15.

Таким образом, вероятность того, что случайная величина X принимает значение в промежутке $[a; b]$, равна интегралу от плотности по этому промежутку (см. рис. 14, а).

Свойства 1° и 2° называют основными свойствами

плотности. Объясняется это тем, что исходя из непрерывной функции $f(x)$, обладающей этими двумя свойствами, по формуле (5) можно построить функцию $F(x)$, обладающую свойствами 1°—5°, т. е. некоторую функцию распределения.

Таким образом, за основу определения закона распределения непрерывной случайной величины X можно было бы взять не функцию распределения $F(x)$, а плотность $f(x)$, т. е. функцию $f(x)$ со свойствами 1° и 2°.

Из свойства 1° следует, что график плотности $f(x)$ расположен над осью Ox . Свойство 2° утверждает, что площадь фигуры под графиком плотности равна единице (см. рис. 14, б). Отсюда, в частности, заключаем, что $f(x) \rightarrow 0$ при $x \rightarrow \pm\infty$, так как в противном случае интеграл по бесконечному промежутку от $f(x)$ был бы равен ∞ . Заметим, наконец, что свойство 2° является аналогом условия полноты для дискретной случайной величины.

В частном случае плотность $f(x)$ может быть отлична от нуля только в некотором промежутке $[\alpha; \beta]$, а вне его равна нулю (рис. 15). Тогда если $a, b < \alpha$ или $a, b > \beta$, то в силу (7) получаем

$$P(a \leq X \leq b) = \int_a^b f(x) dx = \int_a^b 0 \cdot dx = 0.$$

Это означает, что событие $a \leq X \leq b$ невозможно.

Таким образом, если плотность $f(x)$ отлична от нуля только на промежутке $[\alpha; \beta]$, то вне этого промежутка случайная величина X не принимает значений. Все ее значения лежат внутри $[\alpha; \beta]$. При этом

$$\int_{\alpha}^{\beta} f(x) dx = \int_{-\infty}^{\infty} f(x) dx = 1.$$

Резюмируя, можем сказать, что для того чтобы задать закон распределения непрерывной случайной величины X , нужно задать либо ее функцию распределения $F(x) = P(X < x)$, либо плотность распределения $f(x) = F'(x)$.

2. Числовые характеристики непрерывной случайной величины

Так же, как и для дискретной случайной величины, среднее значение и мера рассеяния непрерывной случайной величины характеризуются математическим ожиданием $M(X)$ и дисперсией $D(X)$.

О п р е д е л е н и е. Математическим ожиданием непрерывной случайной величины X с законом распределения $f(x)$ называется число

$$M(X) = \int_{-\infty}^{\infty} xf(x) dx. \quad (8)$$

Формула (8) является аналогом определения математического ожидания, которое мы дали для дискретной случайной величины. В самом деле, пусть для простоты $f(x)$ равна нулю вне некоторого промежутка $[a; b]$. Точками $a = x_0 < x_1 < \dots < x_n = b$ разобьем этот промежуток на достаточно большое число равных промежутков $\Delta_i = [x_{i-1}; x_i]$. Тогда внутри каждого такого промежутка x меняется мало и можно считать, что $x \simeq \tilde{x}_i$, где \tilde{x}_i — середина промежутка Δ_i . Поэтому

$$\begin{aligned} M(X) &= \int_a^b xf(x) dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} xf(x) dx \simeq \\ &\simeq \sum_{i=1}^n \tilde{x}_i \int_{x_{i-1}}^{x_i} f(x) dx = \sum_{i=1}^n \tilde{x}_i p_i, \end{aligned}$$

где p_i — вероятность того, что случайная величина X принимает значения в промежутке Δ_i , т. е. вероятность того, что $X \simeq \tilde{x}_i$. Таким образом,

$$M(X) \simeq \sum_{i=1}^n \tilde{x}_i p_i,$$

что вполне аналогично определению $M(X)$ для дискретной величины.

Теперь мы можем перейти к дисперсии.

Определение. Дисперсией непрерывной случайной величины X называется математическое ожидание квадрата ее отклонения.

Как видим, по содержанию это определение ничем не отличается от определения дисперсии дискретной случайной величины. Отличие состоит в технике вычисления, но и оно не принципиальное. В соответствии с формулой для математического ожидания (8) мы должны написать

$$D(X) = M(X - MX)^2 = \int_{-\infty}^{\infty} (x - MX)^2 f(x) dx. \quad (9)$$

Равенство (8) можно представить в виде

$$\begin{aligned} D(X) &= \int_{-\infty}^{\infty} [x^2 - 2xM(X) + (M(X))^2] f(x) dx = \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2M(X) \int_{-\infty}^{\infty} x f(x) dx + \\ &\quad + (M(X))^2 \int_{-\infty}^{\infty} f(x) dx. \end{aligned}$$

Пользуясь свойствами плотности и определением математического ожидания, отсюда заключаем, что

$$D(X) = M(X^2) - [M(X)]^2. \quad (10)$$

Таким образом, дисперсия непрерывной случайной величины выражается через математическое ожидание, так же как и дисперсия дискретной величины.

Свойства математического ожидания непрерывной случайной величины повторяют соответствующие свойства среднего значения дискретной величины. Отсюда и из формулы (10) следует, что на дисперсию непрерывной величины распространяются свойства дисперсии дискретной величины.

Среднее квадратичное отклонение $\sigma(X)$, как и для дискретной величины, определяется равенством: $\sigma(X) = \sqrt{D(X)}$.

3. Неравенство Чебышева

Значение среднего квадратичного отклонения σ как характеристики рассеяния случайной величины хорошо проявляется в неравенстве, установленном П. Л. Чебышевым. Это неравенство справедливо для обоих типов случайных величин. Но особенно наглядно оно для непрерывной величины. Для этого типа мы его и докажем. (Для дискретной величины доказательство аналогично.)

Т е о р е м а. Пусть X — непрерывная случайная величина с математическим ожиданием $M(X) = a$ и средним квадратичным отклонением $\sigma(X) = \sigma$. Тогда для любого $m > 0$ имеет место неравенство

$$P((x - a) \geq m\sigma) \leq \frac{1}{m^2}. \quad (11)$$

В самом деле, разобьем числовую ось $(-\infty, \infty)$ на два множества: S_1 — промежуток, для которого $|x - a| < m\sigma$, и S_2 — объединение двух бесконечных промежутков, для которых $|x - a| \geq m\sigma$. Тогда можем написать

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x - a)^2 f(x) dx = \\ &= \int_{S_1} (x - a)^2 f(x) dx + \int_{S_2} (x - a)^2 f(x) dx \end{aligned}$$

(здесь знаком \int_{S_i} обозначен определенный интеграл по множеству S_i).

Так как первый интервал неотрицателен, имеем $\sigma^2 \geq \int_{S_2} (x - a)^2 f(x) dx$. На множестве S_2 выполняется неравенство $(x - a)^2 \geq m^2\sigma^2$, и потому

$$\begin{aligned} \sigma^2 &\geq \int_{S_2} (x - a)^2 f(x) dx \geq \int_{S_2} m^2\sigma^2 f(x) dx = \\ &= m^2\sigma^2 \int_{S_2} f(x) dx = m^2\sigma^2 P(X \in S_2). \end{aligned}$$

Отсюда

$$P(X \in S_2) = P(|X - a| \geq m\sigma) \leq \frac{1}{m^2},$$

что и требовалось доказать.

Неравенство (11) называется неравенством Чебышева. Положив $m\sigma = \varepsilon$ и, следовательно, $m = \frac{\varepsilon}{\sigma}$, найдем другую форму этого неравенства:

$$P(|X - a| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}, \quad \text{т. е. } P(|X - a| \geq \varepsilon) \leq \frac{Dx}{\varepsilon^2}. \quad (12)$$

Пользуясь теоремой сложения вероятностей, из (11) и (12) легко получить эквивалентные им неравенства:

$$\begin{aligned} P(|X - a| < m\sigma) &= 1 - P(|X - a| \geq m\sigma) \geq \\ &\geq 1 - \frac{1}{m^2} \end{aligned} \quad (13)$$

и соответственно

$$P(|X - a| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}. \quad (14)$$

Хотя неравенство (11) справедливо для любого $m > 0$, содержательный смысл оно имеет лишь при $m > 1$. Так, например, при $m = 1$ это неравенство утверждает, что вероятность отклонения значений X от среднего значения a больше чем на σ не превосходит 1. Здесь нет никакой новой информации, поскольку вероятность никогда не превосходит 1. Однако, положив $m = 2, 3, 4$ или 5, мы получим новые и очень полезные сведения о любой случайной величине. Оказывается, вероятность отклонения значений любой случайной величины X от среднего значения a больше чем на 2σ не превосходит 0,25, а больше, скажем, на 5σ не превосходит 0,04.

Можно поступить и по-другому — задать вероятность отклонения p и узнать максимальное отклонение, допускаемое этой вероятностью. В самом деле, положим $1 - \frac{1}{m^2} = p$ и, следовательно, $m = \frac{1}{\sqrt{1-p}}$.

Тогда неравенство (13) примет вид

$$P\left(|X - a| < \frac{\sigma}{\sqrt{1-p}}\right) \geq p.$$

Это значит, что с вероятностью, большей чем p , значения любой случайной величины X лежат в промежутке $\left[a - \frac{\sigma}{\sqrt{1-p}}; a + \frac{\sigma}{\sqrt{1-p}} \right]$. Этот промежуток называется *вероятным интервалом* значений. Вероятность p стараются выбирать побольше с тем, чтобы утверждение было «почти» достоверным. Однако при этом вероятный интервал расширяется. Так, при $p = 0,95$ или $0,99$ получаем вероятные интервалы вида $[a - 2\sqrt{5}\sigma; a + 2\sqrt{5}\sigma]$ и соответственно $[a - 10\sigma; a + 10\sigma]$. Таким образом, для любой случайной величины с вероятностью $0,99$ можем утверждать, что все ее значения лежат в промежутке $[a - 10\sigma; a + 10\sigma]$. Вне этого промежутка почти нет значений, точнее не более 1 % значений.

Для конкретных случайных величин с известной плотностью $f(x)$ подобные оценки, как мы увидим в дальнейшем, можно уточнить. Но оценка Чебышева далеко не лишняя. Она указывает вероятность заданного отклонения от среднего всех случайных величин, в том числе и тех, для которых заданы a и σ , но неизвестен закон распределения. Такие величины часто встречаются на практике.

Пример. Предположим, что детали изготавливаются на автоматическом станке и a — проектный размер деталей. Каким бы ни был точным станок, каждая реальная деталь имеет небольшое отклонение от a в ту или иную сторону. Следовательно, размер детали X — случайная величина, и если нет систематических отклонений (станок отрегулирован), то среднее значение $MX = a$.

Точность станка задана его конструкцией. Она указана в паспорте и характеризует допустимые отклонения реальных размеров деталей от проектных. Такая характеристика выражается дисперсией σ^2 или средним квадратичным отклонением σ .

Пользуясь неравенством Чебышева, можно установить вероятность тех или иных отклонений от среднего значения. Предположим, например, что $\sigma = 0,01$ и появилась деталь размером $X = a + 0,11$. Поскольку отклонение $X - a = 0,11 > 10\sigma$, т. е. лежит за пределами 10σ -го отклонения, вероятность обнаруженного значения X не более $0,01$. Если появилось несколько деталей таким размером, то можно с уверенностью

сказать, что станок разрегулирован: либо сместилось среднее значение $MX = a$, либо изменилось σ .

Обычно указывается некоторый допуск ε на размеры. Деталь признается годной, если отклонение $|X - a| \leq \varepsilon$. Из неравенства (12) получаем оценки вероятности изготовления негодной детали при заданных точности σ и допуске ε . Видно, что для того чтобы уменьшить вероятность брака, нужно либо увеличить квадрат допуска, либо уменьшить дисперсию станка.

Разумеется, ситуация со станком и деталями не является какой-то исключительной. Подобные расчеты можно проводить с любой случайной величиной, для которой известны ее средние значения a и среднее квадратичное отклонение σ .

Неравенство Чебышева играет ключевую роль в решении фундаментальной задачи определения среднего значения случайной величины. Мы рассмотрим эту задачу в последней главе.

Нормальное распределение



Практически ни одно серьезное исследование, связанное с обработкой результатов наблюдений, не обходится без распределения, которое связывают с именем гениального немецкого математика и физика К. Ф. Гаусса (1777—1855). Сын поденного рабочего, он уже в 22 года получил докторскую степень за доказательство основной теоремы высшей алгебры о числе корней уравнения n -й степени — доказательство, которое в отличие от имеющихся к тому времени доказательств Д'Аламбера, Эйлера, Лагранжа и Лапласа не содержало ошибок и неточностей.

Занимаясь физикой и астрономией, Гаусс заметил, что если ошибки наблюдений носят случайный характер, то независимо от существа эксперимента они распределяются по строго определенному закону. Этот закон был известен давно и назывался нормальным законом распределения. Отдавая должное заслугам Гаусса в изучении и применении этого закона, его стали называть гауссовым.

Со временем выяснилось, что нормальное распределение посит поистине универсальный характер не только в теории ошибок. Годовое количество осадков в данном районе; размеры взрослой особи данного вида, в частности рост и вес; давление крови и число лейкоцитов у здорового человека; вес и размеры семян; вес новорожденных — эти и многие другие случайные величины распределены, если не точно по нормальному закону, то весьма близко к нему. Наконец, было установлено, что многие теоретические распределения, например распределение Бернулли, при

известных условиях также можно заменить нормальным. Неслучайно поэтому, что нормальное распределение входит в арсенал средств исследования любого грамотного специалиста.

1. Кривая Гаусса

Говорят, что случайная величина X распределена по *нормальному* или *гауссову закону*, если плотность ее распределения имеет вид

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}. \quad (1)$$

Нетрудно убедиться, что эта функция обладает основными свойствами плотности. Она, очевидно, неотрицательна, и интеграл от нее

$$\int_{-\infty}^{\infty} f(x) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-a)^2}{2\sigma^2}} dx = 1$$

в силу равенства $\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}$, установленного Пуассоном.

Каков смысл параметров a и σ , входящих в выражение $f(x)$? Чтобы это выяснить, достаточно найти математическое ожидание и дисперсию нормально распределенной случайной величины. Вычисляя интегралы $MX = \int_{-\infty}^{\infty} xf(x) dx$ и $DX = \int_{-\infty}^{\infty} (x-a)^2 f(x) dx$, обнаружим замечательный факт. Оказывается, что $a = MX$, а $\sigma = \sqrt{DX}$. Таким образом, a — это среднее значение нормально распределенной случайной величины, а σ — ее среднее квадратичное отклонение.

Исследуем теперь график плотности $f(x)$. Его называют *кривой Гаусса*.

Как мы уже установили, $f(x)$ неотрицательна. Кроме того, из (1) непосредственно следует, что $f(x) \rightarrow 0$ при $x \rightarrow \pm\infty$ и $f(x)$ симметрична относительно точки

$x = a$. Найдем точки экстремума функции $f(x)$. Дифференцируя, получим

$$f'(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}} \left(-\frac{1}{2\sigma^2}\right) 2(x-a).$$

Видно, что $f'(x)$ обращается в нуль в единственной точке $x = a$, причем слева от a производная положительна, а справа отрицательна. Значит, $x_0 = a$ — единственная точка экстремума, и этот экстремум — максимум. Подставив $x = a$ в (1), найдем максимальное значение:

$$\max f(x) = f(a) = \frac{1}{\sqrt{2\pi}\sigma}.$$

Аналогично, с помощью второй производной $f''(x)$ можно найти точки перегиба графика $y = f(x)$, отделяющие выпуклые участки от вогнутого:

$$M_1 \left(a - \sigma; \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}} \right) \text{ и } M_2 \left(a + \sigma; \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}} \right).$$

Опираясь на эти сведения, можно представить себе график функции $y = f(x)$. Это колоколообразная кривая с вершиной в точке с абсциссой $x = a$, с горизонтальной асимптотой и двумя симметрично расположенными точками перегиба M_1 и M_2 (рис. 16). Изменение a не меняет форму кривой, а только сдвигает ее — вправо при росте a и влево в противном случае. Что же касается второго параметра σ , то, чем меньше σ , тем выше и уже холм кривой. Это понятно. Ведь σ — мера рассеяния. Чем меньше σ , тем плотнее значения случайной величины группируются вокруг ее среднего значения, т. е. тем больше вероятность таких значений, которые близки к среднему.

Случайную величину с нормальным законом распределения будем кратко называть *нормальной случайной величиной*. В частном случае при $a = 0$ и $\sigma = 1$ нормальный закон

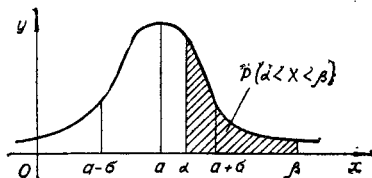


Рис. 16.

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

называется *стандартным*. Очевидно, что плотность стандартного нормального распределения $\varphi(x)$ симметрична относительно начала координат.

2. Правило трех сигм

Пусть X — нормальная случайная величина с параметрами a и σ . Тогда, как известно, вероятность того, что X принимает значения из промежутка $[\alpha; \beta]$, определяется равенством (см. рис. 16):

$$P(\alpha \leq X \leq \beta) = \int_{\alpha}^{\beta} f(x) dx = \frac{1}{\sigma \sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{(x-a)^2}{2\sigma^2}} dx.$$

Интеграл, написанный справа, «неберущийся». Его можно вычислить приближенно для конкретных значений α , β , a и σ . Но делать это специально для каждой новой четверки значений параметров, конечно, нецелесообразно. Указанный интеграл можно преобразовать также к неберущемуся интегралу, но содержащему всего один параметр. Для этого нужно сделать замену переменной $x - a = \sigma t$. Тогда получим

$$P(\alpha \leq X \leq \beta) = \frac{1}{\sqrt{2\pi}} \int_0^{\frac{\beta-a}{\sigma}} e^{-\frac{t^2}{2}} dt - \frac{1}{\sqrt{2\pi}} \int_0^{\frac{\alpha-a}{\sigma}} e^{-\frac{t^2}{2}} dt.$$

Видно, что если мы введем функцию

$$\Phi(x) = \int_0^x \varphi(t) dt = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt, \quad (2)$$

то искомая вероятность выразится как разность ее значений:

$$P(\alpha \leq X \leq \beta) = \Phi\left(\frac{\beta-a}{\sigma}\right) - \Phi\left(\frac{\alpha-a}{\sigma}\right). \quad (3)$$

Функция $\Phi(x)$ называется *интегралом вероятности*, или *функцией Лапласа*. Из определения (2) следует, что $\Phi(x)$ равна вероятности попадания стандартной нормальной величины в промежуток $[0; x]$

(рис. 17). Из той же формулы (2) видно, что $\Phi(x)$ — монотонно возрастающая нечетная функция:

$$\Phi(-x) = -\Phi(x).$$

Составлены достаточно подробные таблицы значений этой функции. Пользуясь ими и формулой (3), нетрудно вычислить вероятность $P(\alpha \leq X \leq \beta)$ для заданной нормальной случайной величины с параметрами a, σ и промежутка $[\alpha; \beta]$. Для этого нужно найти по таблицам значения $\Phi(x)$ в двух точках

$$x_2 = \frac{\beta - a}{\sigma} \text{ и } x_1 = \frac{\alpha - a}{\sigma} \text{ и взять их разность.}$$

В частном случае, если отрезок $[\alpha; \beta]$ симметричен относительно a , т. е. если $\alpha = a - \varepsilon$, $\beta = a + \varepsilon$, $\varepsilon > 0$, то из формулы (3) получаем

$$\begin{aligned} P(a - \varepsilon \leq X \leq a + \varepsilon) &= P(|X - a| \leq \varepsilon) = \\ &= \Phi\left(\frac{\varepsilon}{\sigma}\right) - \Phi\left(-\frac{\varepsilon}{\sigma}\right) = 2\Phi\left(\frac{\varepsilon}{\sigma}\right). \end{aligned}$$

Формула (4) (как и (3)) часто применяется на практике. С ее помощью вычисляется вероятность того, что случайная величина X уклонилась от своего математического ожидания не больше, чем на ε . Так как $\Phi(x)$ растет с ростом аргумента, то из (4) видно, что вероятность попасть в ε -окрестность a тем больше, чем меньше σ . Это вполне согласуется со смыслом σ и характером нормального закона.

В качестве ε — величины отклонения от математического ожидания a — берут обычно число, кратное σ : $\varepsilon = \sigma, 2\sigma, 3\sigma$. Вычислив по таблице соответствующие значения Φ , из (4) получим

$$\begin{aligned} P(|X - a| < \sigma) &= 2\Phi(1) = 0,683, \\ P(|X - a| < 2\sigma) &= 2\Phi(2) = 0,954, \\ P(|X - a| < 3\sigma) &= 2\Phi(3) = 0,997. \end{aligned}$$

Эти равенства наглядно иллюстрируют особенности нормального закона. Первое из них утверждает, что значения нормальной случайной величины с вероят-

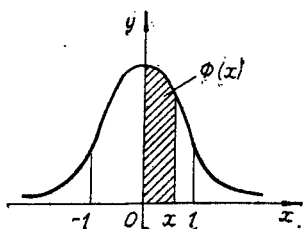


Рис. 17.

ностью, приблизительно равной 0,7, отклоняются от a не более чем на σ . Аналогично, из второго равенства следует, что вероятность отклонения не более чем на 2σ равна 0,954. Это уже близко к 1. Наконец, третье равенство говорит о том, что «почти все» значения лежат в промежутке от $a - 3\sigma$ до $a + 3\sigma$. Иными словами, вероятность того, что нормальная случайная величина принимает значения вне промежутка $[a - 3\sigma; a + 3\sigma]$, ничтожно мала: $1 - P(|X - a| \leq 3\sigma) = 1 - 0,997 = 0,003$. Это значит, что площадь под кривой Гаусса на промежутке $[a - 3\sigma; a + 3\sigma]$ почти равна площади под всей кривой.

Полученный вывод называется *правилом трех сигм*. Оно часто употребляется в различных задачах, связанных с нормальным распределением. В частности, если при испытании выясняется, что случайная величина принимает значения за пределами 3σ от ее математического ожидания, то можно с уверенностью сказать, что либо параметры a и σ вычислены неверно, либо она вообще распределена по какому-то другому закону, отличному от нормального.

Вместо правила трех сигм можно воспользоваться интервалами значений для обычно используемых вероятностей $p = 0,95$ и $0,99$. Построим эти интервалы. Для этого при заданном p найдем столь большое ε , чтобы вероятность попадания значений X в промежуток $[a - \varepsilon; a + \varepsilon]$ была равна p . Пользуясь равенством (4), можем написать

$$\begin{aligned} P(a - \varepsilon \leq X \leq a + \varepsilon) &= P(|X - a| \leq \varepsilon) = \\ &= 2\Phi\left(\frac{\varepsilon}{\sigma}\right) = p. \end{aligned}$$

Отсюда $\Phi\left(\frac{\varepsilon}{\sigma}\right) = \frac{p}{2}$. При $p = 0,95$ имеем $\Phi\left(\frac{\varepsilon}{\sigma}\right) = 0,475$. По таблице значений функции $\Phi(x)$ находим значение аргумента ε/σ , при котором выполняется написанное равенство. С небольшим округлением имеем $\varepsilon/\sigma = 2$, и, следовательно, $\varepsilon = 2\sigma$. Таким образом, с вероятностью $p = 0,95$ можем ожидать значения X в промежутке $[a - 2\sigma; a + 2\sigma]$.

Действуя аналогичным образом, для $p = 0,99$ получим вероятный промежуток $[a - 2,6\sigma; a + 2,6\sigma]$.

Напомним, что в общем случае неравенство Чебышева позволило сделать подобный вывод лишь для значительно более широких промежутков $[a - 2\sqrt{5}\sigma; a + 2\sqrt{5}\sigma]$ и $[a - 10\sigma; a + 10\sigma]$.

3. Отбор деталей и особей

Рассмотрим сначала все тот же пример с изготовлением деталей на автоматическом станке. Что влияет на отклонения размеров деталей X от среднего значения a ? Причин, как правило, очень много: механические колебания фундамента цеха и станка, непостоянство температуры, влажности и давления, небольшие изменения силы или напряжения тока, неоднородность заготовок и т. д. Исключить такие факторы нельзя, по, к счастью, каждый из них оказывает незначительное влияние на изменение X . Многолетний опыт, а затем и теория показали, что в подобных ситуациях случайная величина X распределена по закону, близкому к нормальному. Воспользуемся этим замечательным фактом и решим несколько интересных задач.

Пусть σ — среднее квадратичное отклонение (характеристика точности станка) и ε — установленный допуск на размер. Какова вероятность брака, т. е. какова вероятность неравенства $|X - a| > \varepsilon$?

Пользуясь теоремой сложения вероятностей и формулой (4), можем написать

$$\begin{aligned} P_{\text{бр}} &= P(|X - a| > \varepsilon) = 1 - P(|X - a| \leq \varepsilon) = \\ &= 1 - 2\Phi\left(\frac{\varepsilon}{\sigma}\right). \end{aligned} \quad (5)$$

Эта формула играет ту же роль, что и неравенство Чебышева (12), установленное для произвольной случайной величины. Однако для малых ε , которые как раз и важны при точном производстве, неравенство Чебышева может оказаться вообще неприменимым. В то же время формула (5) — следствие предположения о нормальности распределения X — дает определенный ответ.

Пусть, например, $\sigma = 0,1$ и $\varepsilon = 0,1$. Тогда из неравенства Чебышева (12) получим

$$P_{\text{бр}} = P(|X - a| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} = 1.$$

В этом результате нет содержательной информации, так как всякая вероятность меньше 1. В то же время, вычисляя вероятность брака по формуле (5), будем иметь

$$P_{бр} = 1 - 2\Phi\left(\frac{\varepsilon}{\sigma}\right) = 1 - 2\Phi(1) = 1 - 2 \cdot 0,34 \approx 0,32.$$

Это значение существенно меньше 1, но и не так мал. Во всяком случае такой результат может служить предметом обсуждения. Если будет признано, что вероятность брака слишком велика, то формула (5) подскажет, как эту вероятность уменьшить. Поскольку функция $\Phi(x)$ возрастающая, то, чтобы уменьшить $P_{бр}$, нужно увеличить аргумент ε/σ , т. е. либо увеличить допуск ε , либо отрегулировать станок и уменьшить σ .

Можно поступить и по-другому. Заранее задать требуемую вероятность брака. Например, условиться, что бракованными могут быть не более 5 % деталей. Тогда из неравенства

$$P_{бр} = 1 - \Phi\left(\frac{\varepsilon}{\sigma}\right) < 0,05$$

получим

$$\Phi\left(\frac{\varepsilon}{\sigma}\right) > \frac{0,95}{2} = 0,475.$$

Как мы уже установили, это неравенство выполняется при $\frac{\varepsilon}{\sigma} > 2$, т. е. при $\varepsilon > 2\sigma$. Таким образом, если допуск в 2 раза превосходит среднее квадратичное отклонение, то не более 5 % деталей окажутся бракованными.

Аналогично, ужесточив требование и условившись, что бракованными могут быть не более 1 % деталей, получим $\Phi\left(\frac{\varepsilon}{\sigma}\right) > \frac{0,99}{2} = 0,495$. Отсюда и из таблицы значений $\Phi(x)$ найдем $\varepsilon > 2,6\sigma$.

Еще одна задача в этом же круге. Предположим, что $a = 10$, $\sigma = 0,1$ и для стыковки одной детали с другой требуются размеры от 10,1 до 10,5. Какова

вероятность получения таких деталей на нашем станке? По формуле (3) находим

$$P(10,1 < X < 10,5) = \Phi\left(\frac{10,5 - 10}{\sigma}\right) - \Phi\left(\frac{10,1 - 10}{\sigma}\right) = \\ = \Phi(5) - \Phi(1) = 0,4999 - 0,3413 = 0,1586.$$

Как видим, вероятность не слишком велика. Большая часть деталей не подойдет для стыковки.

Предположим, наконец, что появилось несколько деталей с размерами 10,31 и более. Можно ли считать станок отрегулированным? Так как $10,31 = 10 + 0,31 > a + 3\sigma = 10,3$, можно утверждать, что технология изготовления нарушена — несправлен станок, пришли нестандартные заготовки и т. п.

Примерно такие же проблемы возникают при «производстве» особей потомства в большой популяции. Как мы уже отмечали, многие количественные признаки — вес, рост, объем головы или грудной клетки, длина хвоста, урожайность, удои и т. п. — распределены по нормальному закону с некоторыми параметрами σ и a . Прежде всего возникает вопрос, какие значения количественного признака X мы можем ожидать при данных a и σ ? Например, какие значения X мы можем ожидать с вероятностью 0,95 и какие — с вероятностью 0,99?

Это вопрос о вероятных интервалах нормально распределенной случайной величины. Как мы установили, вероятности $p = 0,95$ соответствуют промежуток $[a - 2\sigma; a + 2\sigma]$, а вероятности $p = 0,99$ — промежуток $[a - 2,6\sigma; a + 2,6\sigma]$.

Часто более приспособленными являются особи, чьи размеры X близки к среднему $MX = a$. Например, хорошо известна связь между весом новорожденного и его выживаемостью. Отклонения как в ту, так и в другую сторону повышают смертность. Если $[a - \varepsilon; a + \varepsilon]$ — промежуток благоприятных, в смысле приспособленности, размеров, то вероятность $P(|X - a| \leq \varepsilon) = 2\Phi(\varepsilon/\sigma)$ (умножения на 100) даст процент приспособленных особей в данной популяции. Ясно, что эта вероятность тем больше, чем больше «допуск» ε и чем меньше среднее квадратичное отклонение σ .

Процент выживших можно определить экспериментально. Например, детская смертность во всех странах

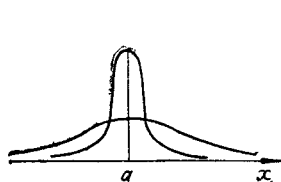


Рис. 18.

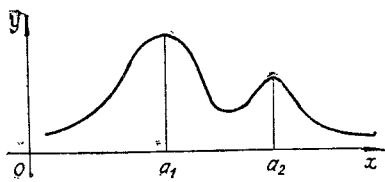


Рис. 20.

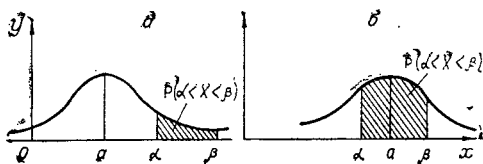


Рис. 19.

изучается самым тщательным образом. Зная вероятность выживания p , из равенства $2\Phi(\varepsilon/\sigma) = p$, как и ранее по таблицам, сумеем найти соответствующее отношение ε/σ . При известном σ это дает возможность определить ширину промежутка благоприятных для выживания размеров.

Особь, чьи размеры не попадают в благоприятный промежуток, в большинстве погибают и, следовательно, не производят потомства. Если признак, о котором идет речь, передается по наследству, то потомство выживших особей также в основном будет обладать благоприятными размерами, т. е. размерами, близкими к среднему. Этот процесс естественного отбора, идущий от поколения к поколению, не изменяет среднего значения a , но уменьшает рассеяние, т. е. уменьшает σ . Холм гауссовой кривой становится более узким и высоким (рис. 18). Такой тип естественного отбора называется стабилизирующим, или центростремительным. Эффект подобного отбора аналогичен регулировке станка, увеличивающей его точность.

С промысловой точки зрения благоприятствующим может оказаться некоторый несимметричный относительно a промежуток $[\alpha; \beta]$. Например, требуются животные определенной упитанности, плоды определенной длины и т. п. В этом случае вероятность

$$P(\alpha \leq X \leq \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right)$$

дает возможность найти процент ценных в промысловом отношении особей (рис. 19).

Понятно, что эта вероятность будет невелика, если промежуток $[\alpha; \beta]$ далеко отстоит от a . Крупные плоды встречаются реже, чем средние. Поэтому если α и β заданы, то стараются с помощью искусственного отбора (селекции) сдвинуть среднее значение a таким образом, чтобы приблизить его к промежутку $[\alpha; \beta]$ или даже включить в него (рис. 19). Иногда такой сдвигающий среднее значение отбор наблюдается и в природе. В этом случае более приспособленными оказываются особи с новым средним значением. Отбор такого типа называют направленным.

Отметим между прочим, что популяция может состоять из двух частей, каждая из которых имеет нормальное распределение, но со своим средним значением и своей дисперсией. Например, популяция больных и здоровых с различным кровяным давлением; популяция нормальных и мутантных особей; популяция, обитающая в двух различных нишах; популяция двух пород животных и т. д. Кривая плотности распределения признака X у таких популяций имеет два холма (рис. 20). Ее называют бимодулярной. Отбор в такой популяции может приводить и к расхождению средних значений (дизруптивный отбор), и к уменьшению дисперсий.

Сдвиг среднего значения может произойти и при эксплуатации популяции. Многие популяции в стационарных условиях стабилизируют свою численность X , после чего X меняется случайным образом по нормальному закону с некоторыми a и σ . Эксплуатация популяции, состоящая в том, что из нее изымается некоторое количество особей, сохраняет нормальный закон изменения численности X , но приводит к резкому уменьшению среднего значения MX . Возникает опасность, что, изменяясь случайным образом с новым средним значением $a_1 < a$, численность X в какой-то момент станет меньше некоторого критического значения β , что приведет к гибели популяции. Вероятность такой опасности $P(0 < X < \beta) = \Phi\left(\frac{\beta - a_1}{\sigma}\right) - \Phi\left(-\frac{a_1}{\sigma}\right)$.



Задолго до начала развития теории вероятностей было замечено, что сумма большого числа случайных величин ведет себя намного проще, чем отдельные слагаемые. Случайные причины, накладываясь одна на другую, часто погашают противоположные отклонения, сглаживают резкие колебания отдельных «солистов». Этот фундаментальный факт впоследствии лег в основу большого круга теорем о предельном поведении суммы случайных величин. Было установлено, что закон распределения суммы большого числа случайных величин достаточно прост. Точнее говоря, его приближенно можно заменить одним из известных распределений. Во многих случаях таким предельным распределением является нормальное распределение Гаусса.

1. Муавр — Лаплас — Ляпунов

С одной предельной теоремой мы уже знакомы. Рассматривая биномиальное распределение, мы доказали теорему Пуассона, из которой следовало, что при малом p и большом n биномиальное распределение $P_n(k)$ может быть приближенно заменено распределением Пуассона $P_k = \frac{a^k}{k!} e^{-a}$. Напомним, что случайная величина X , распределенная по биномиальному закону, представляет собой сумму n одинаково распределенных случайных величин X_i , принимающих всего два значения 1 и 0 с вероятностями соответственно p и q . Таким образом, теорема Пуассона — это теорема о

предельном поведении при $n \rightarrow \infty$ суммы $X = \sum_{i=1}^n X_i$.

Предельная теорема Пуассона была опубликована в начале 30-х годов прошлого века и касалась специального случая биномиального распределения — случая малого p . За 100 лет до этого французский математик А. де Муавр (правда, большую часть своей жизни он прожил в Англии) доказал предельную теорему для биномиального распределения в другом частном случае $p = q = 1/2$. Выяснилось, что в этом случае предельное распределение является нормальным. В 1783 году Лаплас обобщил этот вывод на случай произвольного p .

К концу прошлого века существенно изменились представления о роли теории вероятностей, раздвинулись границы ее приложений. В 1901 году знаменитый русский математик А. М. Ляпунов доказал теорему, которую в силу ее универсальности стали называть *центральной предельной теоремой*. Оказалось, что вывод о «почти нормальном» распределении суммы $\sum_{i=1}^n X_i$ большого числа случайных величин справедлив не только для величин схемы Бернулли, но и для случайных величин X_i любой природы. Требуется только, чтобы X_i были взаимно независимы и вклад каждой из них в сумму не выделялся среди остальных.

Чтобы точно сформулировать теорему Ляпунова, нам придется ввести некоторые обозначения. Пусть X_1, X_2, \dots, X_n — взаимно независимые случайные величины и $X = \sum_{i=1}^n X_i$. Обозначим, как обычно, $M(X)$ и $\sigma(X)$ математическое ожидание и среднее квадратичное отклонение X . Рассмотрим также $c_i^3 = M|X_i - M(X_i)|^3$ и $C^3 = \sum_{i=1}^n c_i^3$. Вместо величины X удобно ввести так называемую *нормированную* величину $Z = \frac{X - M(X)}{\sigma(X)}$. Нетрудно доказать, что $M(Z) = 0$ и $\sigma(Z) = 1$.

Так как $X, M(X)$ и $\sigma(X)$ зависят от n , то и Z зависит от n .

Теорема Ляпунова. Если все c_i ограничены и $\frac{C}{\sigma(X)} \rightarrow 0$ при $n \rightarrow \infty$, то вероятность $P(Z < x)$ стре-

мится к функции распределения стандартной нормальной величины, и, следовательно, при большом n

$$P(\alpha < Z < \beta) \simeq \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{x^2}{2}} dx. \quad (1)$$

В теории условия теоремы Ляпунова обычно проверяются без особого труда. В практических задачах условие «примерно равного» вклада случайных величин X_i в сумму $\sum_{i=1}^n X_i$ следует из физического или биологического существа дела. Например, при изготовлении деталей на автоматическом станке или созревании зерен в поле на размеры деталей и зерен X влияет громадное число факторов — ежедневное и даже ежесекундное изменение влажности, температуры и давления; наличие примесей; упругие и электрические свойства станка, детали и даже инструмента для измерения X ; движение воздуха; солнечная активность; колебания почвы и т. д. и т. п. Каждое из этих воздействий в отдельности мало заметно, а число их очень велико. В таких ситуациях выполняются условия теоремы Ляпунова, и можно считать, что размеры X — результат воздействия суммы большого числа случайных факторов — распределены по нормальному закону или закону, близкому к нормальному. Результаты многочисленных экспериментов подтверждают этот вывод. Для подобных случайных величин, пользуясь нормальным распределением, можно вычислять вероятность значений из данного интервала или, наоборот, определять вероятный интервал значений по заданной вероятности, применять правило трех сигм и т. п. Все это мы проделали в рассмотренных примерах с деталями и особями.

К сожалению, доказательство теоремы Ляпунова очень сложное, и мы его не будем приводить. Что же касается теорем Муавра — Лапласа, то их оригинальные доказательства также непросты. Однако если воспользоваться теоремой Лапласа, то теорема Муавра — Лапласа получается из нее как следствие.

Пусть $X = \sum_{i=1}^n X_i$ — случайная величина схемы Бернулли в серии из n испытаний и $k = 0, 1, 2, \dots$

..., n — ее значения. Предположим, что p не меняется с ростом n . Тогда, как мы знаем, $M(X_i) = p$, $M(X) = np$, $\sigma(X) = \sqrt{npq}$. Поэтому

$$\begin{aligned} c_i^3 &= M|X_i - M(X_i)|^3 = M|X_i - p|^3 = \\ &= |1 - p|^3 p + |0 - p|^3 q = \\ &= (1 - p)^3 p + p^3 (1 - p) = pq(p^2 + q^2). \end{aligned}$$

Видно, что c_i не зависит от i и, следовательно, все c_i ограничены одним и тем же числом. Будем писать вместо c_i просто c . Тогда

$$C = \sqrt[3]{\sum_{i=1}^n c_i^3} = \sqrt[3]{c^3 n} = c \sqrt[3]{n}$$

и
$$\frac{C}{\sigma(X)} \frac{c \sqrt[3]{n}}{\sqrt{npq}} = \frac{c}{\sqrt{pq}} n^{-\frac{1}{6}} \rightarrow 0 \quad \text{при } n \rightarrow \infty.$$

Таким образом, для случайной величины X схемы Бернулли условия теоремы Ляпунова выполняются и, следовательно, при больших n биномиальное распределение близко к нормальному. Точнее говоря, имеет место

Теорема (интегральная теорема Муавра — Лапласа). Если X — число наступлений события A в схеме Бернулли и $Z = \frac{X - M(X)}{\sigma(X)} = \frac{X - np}{\sqrt{npq}}$, то для любых α и β

$$\lim_{n \rightarrow \infty} P(\alpha < Z < \beta) = \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{x^2}{2}} dx.$$

Таким образом, при достаточно большом n

$$P\left(\alpha \leq \frac{X - np}{\sqrt{npq}} \leq \beta\right) \simeq \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{x^2}{2}} dx. \quad (2)$$

На практике вместо равенства (2) удобнее пользоваться равенством для вероятности исходной случайной величины X . Событие $\alpha \leq \frac{X - np}{\sqrt{npq}} < \beta$ эквивалентно событию $np + \alpha\sqrt{npq} \leq X < np + \beta\sqrt{npq}$. Поэтому,

обозначив $k_1 = np + \alpha \sqrt{npq}$, $k_2 = np + \beta \sqrt{npq}$, формулу (2) можем переписать в виде

$$P(k_1 \leq X < k_2) \simeq \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{x^2}{2}} dx, \quad (3)$$

или, пользуясь функцией Лапласа,—

$$P(k_1 \leq X < k_2) \simeq \Phi(\beta) - \Phi(\alpha), \quad (4)$$

где

$$\alpha = \frac{k_1 - np}{\sqrt{npq}}, \quad \beta = \frac{k_2 - np}{\sqrt{npq}}.$$

Равенство (4) играет исключительную роль в приложениях схемы Бернулли. Как мы отмечали, вычисление по формуле Бернулли вероятностей $P_n(k)$ и тем более сумм таких вероятностей — практически невыполнимая задача. Формула (4) сводит решение подобных задач к нахождению по таблицам двух значений функции $\Phi(x)$.

Например, редкое заболевание (или редкий признак) встречается у 1 % особей. Какова вероятность того, что в популяции, содержащей 1000 особей, это заболевание (признак) будет обнаружено не более чем у 20 особей?

Имеем $n = 1000$; $p = 0,01$; $q = 0,99$; $np = 10$; $\sqrt{npq} = 3,146$; $k_1 = 0$, $k_2 = 20$,

$$\alpha = (k_1 - np) / \sqrt{npq} = -10/3,146 = -3,2,$$

$$\beta = (k_2 - np) / \sqrt{npq} = 10/3,146 = 3,2.$$

Поэтому по формуле (4)

$$\begin{aligned} P(0 \leq X \leq 20) &= \Phi(3,2) - \Phi(-3,2) = 2\Phi(3,2) \simeq \\ &\simeq 2 \cdot 0,499 = 0,9986. \end{aligned}$$

Как видим, вероятность $P(0 \leq X \leq 20)$ практически равна единице. Как объяснить этот факт? Может быть, мы взяли слишком широкий диапазон изменения X ? Возьмем для сравнения соседний интервал той же длины. Найдём вероятность $P(20 \leq X \leq 40)$. В этом случае

$$\alpha = \frac{10}{3,146} = 3,2, \text{ а } \beta = \frac{30}{3,146} = 9,6.$$

Поэтому по формуле (4)

$$P(20 \leq X \leq 40) = \Phi(9,6) - \Phi(3,2) = \\ = 0,5 - 0,4993 = 0,0007.$$

Мы получили как будто парадоксальный результат. Два соседних интервала одинаковой длины, но вероятность попасть в один из них практически равна единице, а в другой — нулю!

Парадокс легко объясняется, если мы вспомним о том, что $np = M(X) = a$ и $\sqrt{npq} = \sigma(X) = \sigma$, где a и σ — параметры нормального распределения. В нашем случае $a = np = 10$, $\sigma = \sqrt{npq} = 3,146$.

Интервал $[0; 20]$ накрывает трехсигмовый интервал $[a - 3\sigma; a + 3\sigma]$. Поэтому вероятность попасть в него практически равна единице. Соседний же интервал, хотя и имеет ту же длину, но расположен за пределами трехсигмовой зоны, где случайная величина, распределенная по нормальному закону, почти никогда не принимает значений.

Вернемся к задаче о шахматистах, рассмотренной в разделе, посвященном первому знакомству со схемой Бернулли. Предположим, что матч достаточно длинный и прекращается, когда в четном числе партий кто-нибудь из соперников выиграет не менее половины партий. Оценим вероятность окончания матча при условии примерного равенства квалификации соперников.

Имеем $n = 2N$, $p = q = \frac{1}{2}$, $np = N$, $\sqrt{npq} = \frac{\sqrt{N}}{\sqrt{2}}$, $k_1 = N$, $k_2 = N + 1$,

$$\alpha = \frac{k_1 - np}{\sqrt{npq}} = \frac{N - N}{\sqrt{npq}} = 0, \quad \beta = \frac{k_2 - np}{\sqrt{npq}} = \frac{\sqrt{2}}{\sqrt{N}}.$$

Поэтому по формуле (5)

$$P(N \leq X \leq N + 1) \simeq \frac{1}{\sqrt{2\pi}} \int_0^{\frac{\sqrt{2}}{\sqrt{N}}} e^{-\frac{x^2}{2}} dx.$$

При большом N величина $\frac{\sqrt{2}}{\sqrt{N}}$ достаточно мала. Интеграл по малому промежутку приближенно равен значению подынтегральной функции в начале проме-

жутка, помноженному на длину промежутка. Поэтому можем написать

$$P(N \leq X \leq N+1) = \frac{1}{\sqrt{2\pi}} \cdot e^0 \cdot \frac{\sqrt{2}}{\sqrt{N}} = \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\sqrt{N}}.$$

Видно, что с ростом N вероятность выигрыша половины партий стремится к нулю. Отсюда следует, что для шахматистов равной силы нецелесообразно проводить слишком длинные или безлимитные матчи, если в зачет идут только победы. В таких матчах придется очень долго ждать, пока один из соперников выиграет половину партий.

2. Оценка параметров биномиального распределения

Формулы для вероятностей конкретных распределений мы, как правило, использовали в двух направлениях — прямом, когда по заданным параметрам распределения вычисляется вероятность, и обратном, когда по заданной вероятности вычисляются параметры. Предельные теоремы, связывающие биномиальное распределение с нормальным, дают возможность по заданной вероятности оценить параметры биномиального распределения. Рассмотрим две задачи из этой области, важные с практической точки зрения. В одной из них оценивается n , в другой — p .

Пусть X — случайная величина схемы Бернулли с параметрами p и n . Здесь p , как обычно, — вероятность успеха в одном испытании, а n — число испытаний. Напомним, что среднее значение $MX = a = np$, а дисперсия $\sigma^2 = npq$.

Вместо числа успехов X часто удобно рассматривать относительную частоту успехов X/n или, как говорят, долю успехов. Ее среднее значение, очевидно, равно $M\left(\frac{X}{n}\right) = \frac{1}{n} \cdot MX = p$. В конкретных сериях испытаний реальные значения доли успехов $0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{k}{n}, \dots, 1$, вообще говоря, не совпадают со средним значением p . Но можно надеяться, что, когда n велико, они часто попадают в сравнительно узкую окрестность среднего значения p . Нельзя ли найти

столь большое n , чтобы с заданной вероятностью 0,95 или 0,99 гарантировать отклонение доли успехов от среднего значения не больше, чем на заданную величину ε ? Другими словами, каким должно быть n , чтобы при заданном ε выполнялось неравенство $P\left(\left|\frac{X}{n} - p\right| \leq \varepsilon\right) \geq 0,95$?

Неравенство $\left|\frac{X}{n} - p\right| \leq \varepsilon$ эквивалентно неравенству $\left|\frac{X - np}{\sqrt{npq}}\right| \leq \frac{\varepsilon \sqrt{n}}{\sqrt{pq}}$. Следовательно, надо найти такое n , чтобы выполнялось неравенство

$$P\left(\left|\frac{X - np}{\sqrt{npq}}\right| \leq \frac{\varepsilon \sqrt{n}}{\sqrt{pq}}\right) \geq 0,95. \quad (5)$$

При большом n распределение величины $Z = \frac{X - np}{\sqrt{npq}}$ близко к стандартному нормальному, а значения стандартной нормальной величины с вероятностью 0,95 лежат в промежутке $|Z| \leq 2$. Поэтому если выбрать n так, чтобы $\frac{\varepsilon \sqrt{n}}{\sqrt{pq}} = 2$, то неравенство (5) будет выполнено. Отсюда получаем

$$n = \frac{4nq}{\varepsilon^2}. \quad (6)$$

Аналогично решается вопрос с вероятностью 0,99, для которой значения Z лежат в промежутке $|Z| < 2,6$. В этом случае

$$n = \frac{6,76pq}{\varepsilon^2}.$$

Например, в лабораторной популяции плодовых мушек 25 % особей обладают некоторым признаком A . Сколько нужно взять особей, чтобы среди них с вероятностью 0,95 от 20 до 30 % мушек обладали признаком A ? В данном случае $p = 1/4$, $q = 3/4$, $\varepsilon = \frac{30\% - 20\%}{2 \cdot 100\%} = 0,05$. Поэтому по формуле (6)

$$n = \frac{4 \cdot 1 \cdot 3 \cdot 10^4}{4 \cdot 4 \cdot 25} = 300.$$

Оценка n по формуле (6) с успехом используется при анализе общественного мнения. Предположим, что обсуждается некоторый проект — маршрут транспорта, план застройки, кандидат в президенты и т. п. Проект затрагивает интересы большинства населения крупного города или даже страны. Из выступлений прессы, решихей, принятых на собраниях, писем на радио и телевидение складывается впечатление, что примерно 60 % населения поддерживают этот проект. Но так ли это? Можно ли это проверить, не прибегая к референдуму, требующему больших расходов? Рецет прост. Если предположение о 60 %, т. е. о вероятности $p = 0,6$, верно, то по формуле (6) при $\epsilon = 0,05$ найдем n — число людей, среди которых с вероятностью 0,95 должно оказаться от 55 до 65 % выступающих за проект. В данном случае

$$n = \frac{4 \cdot 0,6 \cdot 0,4 \cdot 10^4}{25} = 16 \cdot 24 = 384.$$

Опросив примерно 400 человек и установив, сколько из них поддерживают проект, проверяем, укладывается ли эта доля в интервал от 55 до 65 %. Если укладывается, то, по-видимому, предположение о 60 % верно. Если же нет (вероятность такого события при $p = 0,6$ равна 0,05), то предположение о 60 % надо отвергнуть.

В рассмотренных примерах по заданной вероятности (0,95 или 0,99) и параметру p мы оценили параметр n . Подобным образом неравенство (5) используется и для оценки параметра p . Ведь в реальных задачах значение p далеко не всегда известно. В таких случаях проводят серии по n испытаний, подсчитывают число успехов k в каждой серии и с помощью полученных относительных частот k/n стараются сделать вывод о возможном значении p . Здесь-то и оказывается полезным неравенство (5).

В самом деле, это неравенство выполняется, если n и ϵ связаны формулой (6). Но в данном случае n задано и нужно определить ϵ . Имеем

$$\epsilon^2 = \frac{4pq}{n}. \quad (7)$$

Так как функция $pq = p(1 - p)$ на промежутке $[0; 1]$

не превосходит $1/4$, из (7) получаем

$$\varepsilon \leq \frac{1}{\sqrt{n}}.$$

Отсюда и из неравенства (5) следует, что с вероятностью 0,95 выполняется неравенство

$$\left| \frac{X}{n} - p \right| \leq \frac{1}{\sqrt{n}},$$

т. е. для конкретных значений k случайной величины X с вероятностью 0,95 имеем

$$\frac{k}{n} - \frac{1}{\sqrt{n}} \leq p \leq \frac{k}{n} + \frac{1}{\sqrt{n}}. \quad (8)$$

Подсчитав в данной серии n испытаний число успехов k , мы по неравенству (8) с вероятностью 0,95 найдем возможные границы для значений p .

Промежуток $\left[\frac{k}{n} - \frac{1}{\sqrt{n}}; \frac{k}{n} + \frac{1}{\sqrt{n}} \right]$ называется 95%-ным доверительным интервалом. Его границы зависят от значений k случайной величины, т. е. меняются случайно. При каждом конкретном k , полученном в эксперименте, доверительный интервал с вероятностью 0,95 покрывает неизвестное значение p .

Аналогично получается 99%-ный доверительный интервал $\left[\frac{k}{n} - \frac{1,3}{\sqrt{n}}; \frac{k}{n} + \frac{1,3}{\sqrt{n}} \right]$.

Например, из большой популяции особей (её называют генеральной совокупностью) наугад отобрали 100 особей. Среди этих 100 у 20 обнаружен некий признак A . Что можно сказать о вероятности p обнаружить этот признак у любой наугад взятой особи из генеральной совокупности? В данном случае $n=100$, $k=20$, поэтому в силу (8) с вероятностью 0,95 заключаем, что $1/10 \leq p \leq 3/10$.

Тот факт, что значение p не слишком отличается от полученного в эксперименте значения относительной частоты 0,2, нас не удивил. Об этом мы могли судить, руководствуясь общими соображениями о смысле вероятности. Однако доверительный интервал указывает конкретные границы возможных значений p , а главное — дает 95%-ную гарантию этих границ.

Подобным образом можно поступить при анализе общественного мнения. Если из 400 опрошенных 240 высказались «за», то в силу (8) с вероятностью 0,95

можно утверждать, что доля p выступающих «за» во всем городе или стране лежит внутри промежутка $[0,6 - 0,05; 0,6 + 0,05]$.

Из формулы (8) видно, что при сохранении относительной частоты k/n доверительный интервал будет тем уже, чем больше n . Например, если среди 1000 особей обнаружилось 200 особей с признаком A , то отношение k/n , как и в последнем примере, будет равно 0,2. Однако 95%-ный доверительный интервал $\left[0,2 - \frac{1}{10\sqrt{10}}, 0,2 + \frac{1}{10\sqrt{10}}\right]$ уже, чем $[0,2 - 0,1; 0,2 + 0,1]$.

Оценка среднего значения p с помощью доверительных интервалов оказывается особенно полезной, когда сравниваются две выборки, взятые из одной генеральной совокупности. Это могут быть, например, детали от двух разных станков, плоды с различных полей, конструкции до и после модернизации, больные в контрольной группе и группе, принимающей новое лекарство, и т. п. В таких случаях требуется установить, повлиял ли новый фактор на значение p .

Например, длительному обследованию подвергаются две группы больных. Одна из них, состоящая из 900 человек, лечится традиционным методом, и к концу обследования 200 человек выздоровели. В другой группе, насчитывающей 400 человек, применяется новый метод лечения. К концу обследования в ней выздоровело 140 человек. Является ли новый метод лечения более эффективным по сравнению с традиционным?

Найдем 95%-ные доверительные интервалы для доли p выздоровевших. Для первой группы имеем $\frac{200}{900} - \frac{1}{30} \leq p \leq \frac{200}{900} + \frac{1}{30}$, для второй $\frac{140}{400} - \frac{1}{20} \leq p \leq \frac{140}{400} + \frac{1}{20}$. Таким образом, с вероятностью 0,95 можем утверждать, что применение традиционного метода лечения (ко всем больным, а не только к больным обследуемой группы!) дает вероятность выздоровления, лежащую в интервале $[0,19; 0,26]$. В то же время вероятность выздоровления при использовании нового метода лежит в интервале $[0,3; 0,4]$. Интервалы не пересекаются. Поэтому с вероятностью 0,95 можем утверждать, что новый метод более эффективен, чем традиционный.

Апалогично получаются и 99%-ные доверительные интервалы: $[0,179; 0,265]$ — для традиционного метода и $[0,285; 0,415]$ — для нового.

Подобным образом сравниваются средние доли заболевших (атеросклерозом, раком, туберкулезом, из-венной болезнью) среди курящих и некурящих, среди соблюдающих диету и несоблюдающих, среди жителей двух городов или двух стран, среди особей двух поколений и т. п. Понятно, что вместо болезней можно рассмотреть любой признак у людей и животных, брак у деталей, свойства материалов и др.

* *

*

Так всегда бывает в науке и в жизни. После того как создан сложный и могучий аппарат, потребовавший большого труда и неординарных способностей, конкретные задачи решаются просто. Нам потребовалось изучить биномиальное распределение Бернулли, нормальное распределение Гаусса, использовать предельную теорему Ляпунова. Все это далеко не простые вещи. Но зато теперь мы без труда можем найти, например, доверительный интервал для доли признака. При этом требуется совсем немного информации. Нужно знать общее число наблюдений n и число наблюдений k , в которых данный признак обнаружен. Правда, при этом не нужно забывать, что число наблюдений n должно быть достаточно большим, порядка сотен или тысяч.



Итак, мы установили, что сумма большого числа случайных величин, как правило, имеет сравнительно простое, а именно — нормальное распределение. Однако этим исключительно важным обстоятельством эффект суммы не исчерпывается. На практике часто приходится рассматривать случайную величину \bar{X} , равную сумме n случайных величин X_1, X_2, \dots, X_n , деленной на их число:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Назовем \bar{X} *средней арифметической величиной* величин X . Оказывается, что если n достаточно велико, то при некоторых достаточно общих условиях арифметическая \bar{X} ведет себя «почти» как постоянная величина. Это значит, что дисперсия \bar{X} очень мала, а график плотности распределения \bar{X} представляет собой очень узкий и высокий всплеск около $M\bar{X}$. Первой теоремой подобного рода была теорема Бернулли, доказанная им в 1713 году. Эта теорема получила название «закон больших чисел» и легла в основу создания теории вероятностей. Сейчас законом больших чисел часто называют и другие теоремы о поведении суммы большого числа случайных величин.

Мы получим теорему Бернулли как следствие общей теоремы Чебышева, доказанной им в 1866 году.

1. Теоремы Чебышева и Бернулли

Пусть $X_1, X_2, \dots, X_n, \dots$ — независимые случайные величины с какими-то распределениями и \bar{X} — средняя арифметическая первых n величин:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Разумеется, \bar{X} зависит от n , но мы для краткости вместо $\bar{X}(n)$ будем писать просто \bar{X} .

Теорема Чебышева. Если $X_1, X_2, \dots, X_n, \dots$ — независимые случайные величины, дисперсия которых ограничена одним и тем же числом, то для любого $\varepsilon > 0$ имеет место равенство

$$\lim_{n \rightarrow \infty} P(|\bar{X} - M(\bar{X})| < \varepsilon) = 1. \quad (1)$$

Таким образом, при больших n

$$P(|\bar{X} - M(\bar{X})| < \varepsilon) \simeq 1. \quad (2)$$

Равенство (2) означает, что средняя арифметическая \bar{X} ведет себя «почти» как постоянная. Ее малое отклонение от постоянной $M(\bar{X})$ при больших n — «почти» достоверное событие.

Для доказательства теоремы применим неравенство Чебышева к случайной величине \bar{X} . В силу этого неравенства для любого $\varepsilon > 0$ можем написать

$$P(|\bar{X} - M(\bar{X})| < \varepsilon) \geq 1 - \frac{D(\bar{X})}{\varepsilon^2}. \quad (3)$$

Так как по условию теоремы $D(X_i) < C$ для всех i , то

$$D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) < \frac{C}{n}.$$

Подставив это в неравенство (3), будем иметь

$$P(|\bar{X} - M(\bar{X})| < \varepsilon) \geq 1 - \frac{C}{n\varepsilon^2}.$$

Переходя в этом неравенстве к пределу при $n \rightarrow \infty$, получим требуемое.

В частности, если все X_i имеют одно и то же математическое ожидание $M(X_i) = a$ (этот случай часто встречается на практике), то

$$M(\bar{X}) = \frac{1}{n} \sum_{i=1}^n M(X_i) = \frac{1}{n} \cdot na = a,$$

и в силу (2) —

$$P(|\bar{X} - a| < \varepsilon) \simeq 1, \quad (4)$$

т. е. при большом n с вероятностью, близкой к 1, значения средней арифметической из случайных величин X_1, X_2, \dots, X_n сколь угодно мало отличаются от их общего среднего значения.

Подчеркнем еще раз, что теорема Чебышева, например, в форме (4), не утверждает, что значения \bar{X}

стремятся к a при $n \rightarrow \infty$. Величина $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ — по-прежнему случайная величина, и ее отдельные значения могут быть достаточно далекими от a . Но вероятность таких далеких отклонений с ростом n стремится к нулю. Такое свойство \bar{X} называется *стремлением по вероятности*.

Применим теорему Чебышева к случайным величинам X_i схемы Бернулли. В этом случае, как известно, $M(X_i) = p$, $D(X_i) = pq = p - p^2$. На промежутке $[0; 1]$ значения функции $p - p^2$ ограничены числом $1/4$. Таким образом, $D(X_i) \leq 1/4$ при всех i , условия теоремы Чебышева выполнены, поэтому в силу (4) можем написать

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - p\right| < \varepsilon\right) \simeq 1.$$

Здесь $\sum_{i=1}^n X_i = X$ — число появлений события A в серии из n испытаний. Обозначив k значения X , последнее равенство можем представить в виде

$$P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) \simeq 1. \quad (5)$$

Равенство (5) составляет содержание закона больших чисел Бернулли. Из него следует, что при доста-

точно большом n сколь угодно малые отклонения относительной частоты k/n от p — почти достоверные события, а большие отклонения — почти невозможные события. Полученный вывод об устойчивости относительной частоты оправдывает статистический подход к определению вероятности p как некоего числа, около которого колеблется относительная частота k/n .

Этот вывод вполне согласуется с оценкой p с помощью доверительных интервалов.

2. Несколько слов о математической статистике

Мы рассмотрели всего три закона распределения случайной величины — биномиальный, Пуассона и нормальный. В теории и на практике широко используются и другие законы — геометрический, гипергеометрический, экспоненциальный, равномерный, распределение Стюдента, распределение «хи-квадрат» и другие.

Но как узнать, по какому закону распределена реальная случайная величина X , встретившаяся в данном конкретном эксперименте? Во многих случаях, например в схеме Бернулли или в условиях теоремы Ляпунова, этот закон можно с большей или меньшей степенью правдоподобия обосновать. Иногда характер закона диктуется общими соображениями, вытекающими из существа эксперимента. Но и в этих случаях остаются пока неизвестными параметры распределения X , такие как среднее значение, дисперсия и т. п.

Сам эксперимент дает очень ограниченную информацию о случайной величине. Обычно это некоторый набор ее значений, полученный в результате наблюдений и измерений: x_1, x_2, \dots, x_n . Этот набор называют *выборкой*. Как правило, выборка не исчерпывает всего множества значений X и принципиально не может его исчерпать. Такая ситуация возникает, например, когда для получения значения x_i необходимо уничтожить обследуемый объект (подвергнуть продукт химическому анализу, пробить броню, разрушить кирпич, разрезать плод и т. п.) или когда часть значений может быть получена лишь в будущем, или, наконец, когда число объектов наблюдения слишком велико и их сплошное обследование нецелесообразно экономически.

Так вот, оказывается, что если объем выборки n достаточно велик, то, опираясь на выборку, можно сделать вывод о свойствах самой случайной величины X в целом — о ее математическом ожидании, дисперсии, законе распределения, зависимости от другой величины и т. п. Полученные сведения, хотя и носят вероятностный характер, позволяют приписывать обоснованные решения о качестве партии продукции, о необходимых расходах и запасах, о целесообразности планируемых мероприятий; они дают возможность строить достоверный прогноз. Разработкой способов решения подобных задач занимается бурно прогрессирующая область теории вероятностей — математическая статистика.

Одну из самых простых задач математической статистики мы уже решили. Используя близость при большом n биномиального распределения к нормальному, мы сумели по результатам наблюдений построить доверительные интервалы для параметра p биномиального распределения.

Ключевую роль при решении задач математической статистики играет закон больших чисел. Рассмотрим несколько классических задач.

3. Оценки среднего значения и дисперсии

Пусть целью эксперимента является определение среднего размера каких-то объектов — деталей, плодов, особей и т. п. В таких случаях обычно поступают следующим образом. Предполагают, что размеры объектов — случайная величина X с некоторым средним значением a . Это a и нужно оценить. Для этого из всего множества объектов — генеральной совокупности — выбирают наугад некоторое количество n , снимают искомый размер, получают выборку значений X : x_1, x_2, \dots, x_n и в качестве среднего значения a берут среднее арифметическое значений x_i :

$$a \simeq \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Казалось бы, простая и естественная процедура, знакомая человечеству с незапамятных времен. Тем не менее возникает несколько вопросов. Ведь среднее

арифметическое, вообще говоря, разное при разных выборках; его значение зависит и от n . Какое же из этих значений брать в качестве оценки a ? Почему мы уверены или с большой вероятностью можем утверждать, что любое из этих значений среднего арифметического мало отличается от a ? Каким, наконец, должно быть n , чтобы эту вероятность сделать достаточно большой?

Попробуем ответить на эти вопросы с помощью закона больших чисел.

Будем считать результат i -го измерения x_i значением случайной величины X_i . Все случайные величины x_i распределены так же, как исходная величина X , и имеют, следовательно, те же математическое ожидание a и дисперсию σ^2 . Кроме того, поскольку измерения проводятся независимо одно от другого, все X_i независимы. Образует случайную величину $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$. Очевидно, что ее значениями

являются средние арифметические различных серий измерений $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$, приближенно равные a . Поэтому \bar{X} называют *статистической оценкой* a .

Как распределена случайная величина \bar{X} ? Разумеется, мы не можем подробно описать это распределение, не зная, как распределена X и, следовательно, X_i . Тем не менее некоторые особенности \bar{X} , важные для нашей задачи, мы без труда установим.

Прежде всего, в силу свойств математического ожидания

$$\begin{aligned} M(\bar{X}) &= M\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] = \\ &= \frac{1}{n} \sum_{i=1}^n MX_i = \frac{1}{n} \cdot n \cdot a = a. \end{aligned}$$

Таким образом, математическое ожидание $M\bar{X}$ совпадает со средним значением a исходной величины X . Это означает, что значения \bar{X} с одинаковой частотой лежат и слева и справа от a . Поэтому, выбирая наугад эти значения, мы будем избавлены от систематической ошибки. Оценку, обладающую таким качеством, называют *несмещенной*.

Далее, в силу свойств дисперсии

$$D(\bar{X}) = D\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] = \\ = \frac{1}{n^2} \sum_{i=1}^n DX_i = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

Отсюда видно, что с ростом n рассеяние \bar{X} убывает, ее значения \bar{x} все более и более плотно группируются около среднего значения a . Подставив значение $D(\bar{X})$ в неравенство Чебышева (3), получим

$$P(|\bar{X} - a| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2 n}. \quad (6)$$

Таким образом, для любого сколь угодно малого $\varepsilon > 0$ можно указать столь большой объем выборки n , что вероятность отклонения значений \bar{X} от a меньше чем на ε будет практически равна 1. Это свойство оценки \bar{X} называют *состоятельностью*.

Несмещенность и состоятельность оценки \bar{X} объясняют, почему при большом n вместо неизвестного математического ожидания $MX = a$ можно брать конкретное реализованное значение \bar{x} случайной величины \bar{X} , т. е. среднее значение выборки. Вероятность большого расхождения при этом мала и стремится к нулю при $n \rightarrow \infty$.

Если известна дисперсия σ^2 (например, точность станка) и задана вероятность P малого отклонения \bar{x} от a , т. е. требуется, чтобы

$$P(|\bar{X} - a| < \varepsilon) = p, \quad (7)$$

где p равно, например, 0,95 или 0,99, то можно найти объем выборки n , обеспечивающий эту вероятность. В самом деле, из (6) следует, что указанное требование будет выполнено, если

$$1 - \frac{\sigma^2}{\varepsilon^2 n} = p, \text{ т. е. } n = \frac{\sigma^2}{\varepsilon^2 (1 - p)}. \quad (8)$$

Например, если мы хотим, чтобы с вероятностью $p = 0,99$ случайное значение \bar{x} отклонилось от a не

более чем на 0,1, нам придется провести

$$n = \frac{\sigma^2}{10^{-2} \cdot 10^{-2}} = 10^4 \sigma^2 \quad (9)$$

измерений.

Равенство (9) еще раз подчеркивает роль дисперсии σ^2 как характеристики рассеяния значений случайной величины: чем меньше σ^2 , т. е. чем плотнее расположены значения около среднего значения $MX = a$, тем меньшим числом измерений можно найти приближенное значение a .

Наконец, по заданной вероятности p и числу измерений n можно найти доверительный интервал, который с вероятностью p покрывает неизвестное значение a . В самом деле, из (7) следует, что с вероятностью p выполняется неравенство $|\bar{x} - a| < \varepsilon$, т. е. неравенство $\bar{x} - \varepsilon < a < \bar{x} + \varepsilon$. Подставив сюда значение ε , найденное из (8), получим доверительный интервал для a :

$$\bar{x} - \frac{\sigma}{\sqrt{n} \sqrt{1-p}} < a < \bar{x} + \frac{\sigma}{\sqrt{n} \sqrt{1-p}}.$$

Значения p , σ и n заданы, значение \bar{x} определяется по выборке.

В частности, при $p = 0,95$ и $0,99$ получаем соответственно доверительные интервалы

$$\left[\bar{x} - \frac{2 \sqrt{5} \sigma}{\sqrt{n}}; \bar{x} + \frac{2 \sqrt{5} \sigma}{\sqrt{n}} \right] \text{ и } \left[\bar{x} - \frac{10 \sigma}{\sqrt{n}}; \bar{x} + \frac{10 \sigma}{\sqrt{n}} \right].$$

Мы оценили среднее значение a случайной величины X с помощью несмещенной и состоятельной статистической

оценки $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, которую называют *выборочной средней*. Похожим образом оценивают и дисперсию σ^2 . Статистической оценкой для нее служит случайная величина

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}).$$

Ее называют *исправленной выборочной дисперсией*. Можно показать, что исправленная выборочная дисперсия является также несмещенной и состоятельной оценкой дисперсии σ^2 . Отсюда следует, что при большом n

значения S^2 , реализованные на конкретной выборке, т. е. числа $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, с большой вероятностью располагаются вблизи неизвестной дисперсии σ^2 . Эти числа и берут в качестве приближенного значения σ^2 .

4. Оценки параметров нормального распределения

Используя неравенство и теорему Чебышева, мы построили доверительные интервалы для среднего значения случайной величины X , ничего не предполагая заранее об ее законе распределения. Интервалы довольно широкие. Чтобы их сузить, нужно брать выборки большого объема n . Разумеется, в общем случае и такая информация полезна. Однако если имеются сведения о характере закона распределения X , то можно получить более узкие доверительные интервалы при том же n .

Предположим, например, что X — нормально распределенная случайная величина. Известная ее дисперсия σ^2 , но неизвестно среднее значение a . Построим

статистическую оценку $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, где X_i рас-

пределены так же, как и X . Можно показать, что \bar{X} — тоже нормально распределенная случайная величина. Ее параметры мы уже находили: $M(\bar{X}) = a$.

$D(\bar{X}) = \frac{\sigma^2}{n}$. Следовательно, $\sigma(\bar{X}) = \sqrt{D(\bar{X})} = \frac{\sigma}{\sqrt{n}}$.

Пользуясь нормальностью \bar{X} , можем написать

$$P(|\bar{X} - a| < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sigma(\bar{X})}\right) = 2\Phi\left(\frac{\varepsilon \sqrt{n}}{\sigma}\right).$$

Потребовав, чтобы неравенство $|\bar{X} - a| < \varepsilon$ выполнялось с вероятностью $p = 0,95$, получим $2\Phi\left(\frac{\varepsilon \sqrt{n}}{\sigma}\right) = 0,95$.

Отсюда, как мы это уже не раз делали, по таблице Φ найдем $\frac{\varepsilon \sqrt{n}}{\sigma} = 2$, т. е. $\varepsilon = \frac{2\sigma}{\sqrt{n}}$. Таким образом, с вероятностью 0,95 выполняется неравенство $|\bar{X} - a| <$

$< \frac{2\sigma}{\sqrt{n}}$, или $\bar{x} - \frac{2\sigma}{\sqrt{n}} < a < \bar{x} + \frac{2\sigma}{\sqrt{n}}$. Это и есть

95%-ный доверительный интервал для среднего значения a нормально распределенной случайной величины. Он заметно уже, чем соответствующий интервал в общем случае.

Подобно этому строится и 99%-ный доверительный интервал: $\left[\bar{x} - \frac{2,6\sigma}{\sqrt{n}}; \bar{x} + \frac{2,6\sigma}{\sqrt{n}} \right]$.

Рассмотрим простой пример. Предположим, что изделия производства, образцы эксперимента или природные объекты сортируются в зависимости от среднего значения некоторой величины X . Это может быть линейный размер детали, процент содержания примеси в химическом веществе, диаметр горошины и т. п. Распределение X предполагается нормальным с дисперсией $D(X) = \sigma^2 = 0,04$. Если $M(X) = a = 9 \pm 0,2$, то партии изделий присваивается первый сорт.

В результате обследования взятых наугад 20 изделий из партии получены следующие результаты:

9,1	8,8	9,0	9,4	8,9	8,7	9,1	9,0	8,8	8,9
9,2	9,1	9,0	9,2	8,9	9,0	9,3	8,7	8,9	9,0

Можно ли с вероятностью 0,99 считать, что партия имеет первый сорт?

Сложив все значения x_i и разделив на их число $n = 20$, получим $\bar{x} = 9$. Вычислив $\frac{2,6\sigma}{\sqrt{n}} = \frac{2,6 \cdot 0,2}{\sqrt{20}} = 0,12$, найдем 99%-ный доверительный интервал $[9 - 0,12; 9 + 0,12]$. Он оказался уже, чем допустимый интервал для первосортного среднего размера. Поэтому с вероятностью 0,99 можем утверждать, что вся партия изделий принадлежит первому сорту.

Заметим, что если бы не было известно, что X распределена нормально, нам пришлось бы пользоваться 99%-ным доверительным интервалом общего случая $\left[\bar{x} - \frac{10\sigma}{\sqrt{n}}; \bar{x} + \frac{10\sigma}{\sqrt{n}} \right]$. Он шире допустимого интервала первосортных размеров. Поэтому мы не смогли бы с вероятностью 0,99 рассчитывать на первый сорт. Однако 95%-ный интервал $\left[\bar{x} - \frac{2\sqrt{5}}{\sqrt{n}} \sigma; \bar{x} + \frac{2\sqrt{5}}{\sqrt{n}} \sigma \right]$ совпадает с допустимым интервалом $[9 - 0,2; 9 + 0,2]$.

Следовательно, снизив вероятность до 0,95, можно надеяться на первосортность партии, не опираясь на предложение о нормальности.

До сих пор мы предполагали, что дисперсия σ^2 известна. Вообще говоря, обстоятельства эксперимента могут подсказать лишь нормальный характер распределения, оставляя неизвестными оба параметра: μ и σ . В таких случаях дисперсию σ^2 оценивают с помощью исправленной выборочной дисперсии S^2 . Можно построить и доверительный интервал этой оценки:

$\left(\frac{s^2 (n-1)}{C_1}, \frac{s^2 (n-1)}{C_2} \right)$, где S^2 — значение S^2 на данной выборке, а C_1 и C_2 определяются по специальной таблице, составленной для так называемого распределения «хи-квадрат». Это распределение, как и нормальное, часто используется в теории и на практике.

За границами книги



В небольшой научно-популярной книге, конечно, невозможно даже бегло охватить все разделы теории вероятностей, математической статистики и их приложений в естествознании. Мы и не ставили такой цели. В нашу задачу входило знакомство с первоначальными, наиболее употребляемыми и простыми понятиями и методами. Многие теоремы, о которых шла речь, доказаны 150 и даже 200 лет назад. Это никак не умаляет их значения. Подобно теореме Пифагора, которая намного старше, достижения классиков теории вероятностей навсегда вошли в интеллектуальный фонд человечества и успешно служат ему.

За рамками книги осталось очень многое: теория зависимых случайных величин, опираясь на которую изучают, например, увеличение частоты заболеваний с ростом промышленных загрязнений среды или влияние примесей в металле на частоту появления бракованных изделий; теория стохастических цен; исследующая вероятностные переходы системы, например ракеты, популяции, завода, экономики региона или страны, из одного состояния в другое; теория случайных процессов, т. е. распределений, параметры которых меняются с течением времени; теория временных рядов, т. е. закономерностей, зависящих от времени, на которые накладываются случайные помехи — шумы — этим и другим разделам теории вероятностей посвящены десятки монографий.

Что касается приложений, то во многие науки теория вероятностей проникла так глубоко, что послужила основанием для возникновения новых научных направлений. Таковы статистическая физика, изучающая

с позиций теории вероятностей массовые явления в микромире; теория надежности, устанавливающая вероятный срок службы сложных машин и механизмов в зависимости от параметров составляющих их частей; теория игр, на которую опирается рыночная экономика; теория массового обслуживания, устанавливающая законы функционирования транспорта, больниц, поликлиник, магазинов, мастерских по ремонту и другие.

Успешно развиваются «ведомственные» статистики — социальная, медицинская, экономическая, лингвистическая, сельскохозяйственная и т. п. Особое место занимает биометрия — математическая статистика в биологии. Этот термин 100 лет назад ввел в науку Ф. Гальтон. В 1900 году К. Пирсон основал журнал «Биометрика», который и поныне является одним из авторитетнейших изданий в области математического моделирования в биологии, в частности в биометрии.

Несмотря на различие в источниках задач, которые решаются в таких далеких друг от друга областях, как, например, лингвистика и медицина, их математическая, т. е. собственно статистическая, суть — обработка большого числа экспериментальных данных — часто оказывается совпадающей. Глубоко разработанные универсальные методы математической статистики позволяют по одним и тем же алгоритмам устанавливать автора предъявленного литературного текста и характер болезни по наблюдаемым симптомам.

Дополнительным стимулом успеха статистики в приложениях послужило стремительное развитие вычислительной техники. В современные компьютеры, даже карманные, встроены специальный «статистический блок». Набрав программу этого блока, т. е. нажав в определенной последовательности несколько кнопок, можно затем запускать в этот блок экспериментальные данные — производительность различных заводов; рождаемость, заболеваемость и смертность в различных областях; аварийность и травматизм на различных трассах; урожайность, количество осадков, процент брака, количество вредных примесей, обеспеченность лекарствами и детскими садами, количество книг, кинотеатров, магазинов, издателей и т. д. и т. п. Уже через несколько секунд на экране компьютера появятся математическое ожидание, дисперсия, доверитель-

ный интервал, коэффициент линейной регрессии и все остальное, что было заказано программой.

Нужда в специалистах по теории вероятностей и математической статистике во всех развитых странах неуклонно растет. Спрос на них уступает только спросу на программистов. Эта тенденция объясняется тем, что во всех областях все более важное место отводится изучению массовых явлений и процессов, все более важную роль играют стохастические принципы мышления. Однако подготовка даже большого отряда профессионалов-статистиков не решит проблемы. Речь идет о том, чтобы каждый инженер, врач, квалифицированный рабочий, лингвист и экономист, не говоря уж о биологах, химиках и физиках, вообще, чтобы каждый культурный человек овладел элементами теории вероятностей. Хотелось бы надеяться, что в этом будет полезен предлагаемый список литературы.

1. Вентцель Е. С. Элементы теории игр.— М.: Физматгиз, 1961.
2. Вентцель Е. С. Исследование операций — задачи, принципы, методология.— М.: Наука, 1980.
3. Гнеденко Б. В. Из истории науки о случайном.— М.: Знание, 1981.
4. Гнеденко Б. В., Соловьев А. Д. Математика и теория надежности.— М.: Знание, 1982.
5. Гнеденко Б. В., Коваленко И. И. Введение в теорию массового обслуживания.— М.: Наука, 1987.
6. Глотов Н. В., Животовский Л. А., Хованов П. В., Хромов-Борисов И. И. Биометрия.— Л.: Изд-во ЛГУ, 1982.
7. Гроссман С., Тернер Дж. Математика для биологов.— М.: Высш. шк., 1983.
8. Кендалл М., Моран П. Геометрические вероятности.— М.: Наука, 1972.
9. Кимура М. Молекулярная эволюция: теория нейтральности.— М.: Мир, 1985.
10. Компанец А. С. Законы физической статистики.— М.: Наука, 1976.
11. Ли Ч. Ч. Введение в популяционную генетику.— М.: Мир, 1978.
12. Лакин Г. Ф. Биометрия.— М.: Высш. шк., 1990.
13. Моран П. Статистические процессы эволюционной теории.— М.: Наука, 1973.
14. Прохоров Ю. В., Розанов Ю. А. Теория вероятностей.— М.: Наука, 1987.
15. Проблемы теории молекулярной эволюции/Ратнер В. А., Жарких А. А., Колчанов Н. А. и др.— Новосибирск: Наука. Сиб. отд-ние, 1985.
16. Скороход А. В. Вероятность вокруг нас.— Киев: Наук. думка, 1980.

17. Тарасов Л. В. Мир, построенный на вероятности.— М.: Просвещение, 1984.
18. Урбах В. Ю. Статистический анализ в биологических и медицинских исследованиях.— М.: Медицина, 1975.
19. Феллер В. Введение в теорию вероятностей и ее приложения.— М.: Мир, 1984.— Т. 1, 2.
20. Хургин Я. И. Да, нет или может быть...— М.: Наука, 1983.
21. Хургин Я. И. Как объять необъятное.— М.: Знание, 1985.
22. Чистяков В. П. Курс теории вероятностей.— М.: Наука, 1982.
23. Шредингер Э. Что такое жизнь с точки зрения физика?— М.: Атомиздат, 1972.
24. Эйген М. Самоорганизация материи и эволюции биологических макромолекул.— М.: Мир, 1973.
25. Яглом А. М., Яглем И. М. Вероятность и информация.— М.: Наука, 1973.

Наука об умении угадывать	3
Случайные события	8
1. Испытания, события, вероятности	—
2. Классическая вероятность	12
3. Классическая вероятность в генетике	15
4. Некоторые формулы комбинаторики	22
5. Комбинаторика и генетика	27
6. Геометрическая вероятность	37
7. Статистическая вероятность	45
8. Вероятности большие и маленькие	46
Аксиомы и теоремы вероятности	48
1. Случайные события в общей схеме	49
2. Алгебра событий	51
3. Аксиомы вероятности. Теоремы сложения	54
4. Независимость событий. Условная вероятность	58
Случайные события в биологии	65
1. Вероятность суммы и пересечения признаков	—
2. Теория встреч	68
3. Вероятность цепи	73
4. Дублирование	77
5. Последовательные и параллельные соединения	80
6. Задача о числе дублеров	81
7. Условная и полная вероятность	87
Дискретная случайная величина	103
1. Числовые события	—
2. Закон распределения	104
3. Схема Бернулли. Биномиальное распределение	106
4. Распределение Пуассона	113
Действия со случайными величинами	124
1. Равенство двух случайных величин	—
2. Сумма случайных величин	125
3. Произведение случайных величин	127
Числовые характеристики дискретной случайной величины	129
1. Математическое ожидание	—
2. Свойства математического ожидания	132
3. Дисперсия	135
4. Числовые характеристики биномиального распределения	138
5. Числовые характеристики распределения Пуассона	139
6. Математическое ожидание в генетике	142

Непрерывная случайная величина	149
1. Закон распределения непрерывной случайной величины	149
2. Числовые характеристики непрерывной случайной величины	155
3. Неравенство Чебышева	157
Нормальное распределение	161
1. Кривая Гаусса	162
2. Правило трех сигм	164
3. Отбор деталей и особей	167
Предельные теоремы	172
1. Муавр — Лаплас — Ляпунов	—
2. Оценка параметров биномиального распределения	178
Закон больших чисел	184
1. Теоремы Чебышева и Бернулли	185
2. Несколько слов о математической статистике	187
3. Оценки среднего значения и дисперсии	188
4. Оценки параметров нормального распределения	192
За границами книги	195

Научно-популярное издание

Гильдерман
Юрий Ильич

ЗАКОН И СЛУЧАЙ

Редакторы издательства
Л. В. Филиппова, Л. П. Голышева
Художник С. М. Кудрявцев
Художественный редактор С. В. Марковская
Технический редактор Г. Я. Герасимчук
Корректор Г. И. Шведкина

ИБ № 42744

Сдано в набор 01.08.90. Подписано к печати 08.08.91. Формат 84×108¹/₃₂. Бумага кн.-журнальная. Обыкновенная гарнитура. Высокая печать. Усл. печ. л. 10,5. Усл. кр.-отт. 10,9. Уч.-изд. л. 10. Тираж 10 000 экз. Заказ № 322. Цена 2 руб.

Ордена Трудового Красного Знамени издательство «Наука», Сибирское отделение, 630099 Новосибирск, ул. Советская, 18.
4-я типография издательства «Наука».
630077 Новосибирск, ул. Станиславского, 25.

