ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Что за польза от книги без картинок и разговоров?

Льюис Кэрролл, «Приключения Алисы в стране чудес»

Перед Вами, уважаемый читатель, итог размышлений автора о содержании начального курса математической статистики. Настоящая книга — это, в первую очередь, множество занимательных примеров и задач, собранных из различных источников. Задачи предназначены для активного освоения понятий и развития у читателя навыков квалифицированной статистической обработки данных. Для их решения достаточно знания элементов математического анализа и теории вероятностей (краткие сведения по теории вероятностей и линейной алгебре даны в приложении).

Акцент делается на наглядном представлении материала и его неформальном пояснении. Теоремы, как правило, приводятся без доказательств (со ссылкой на источники, где их можно найти). Наша цель — и осветить практически наиболее важные идеи математической статистики, и познакомить читателя с прикладными методами.

Первая часть книги (гл. 1–5) может служить введением в теорию вероятностей. Особенностью этой части является подход к освоению понятий теории вероятностей через решение ряда задач, относящихся к области статистического моделирования (имитации случайности на компьютере). Ее материал, в основном, доступен школьникам старших классов и студентам 1-го курса.

Вторая и третья части (гл. 6–13) посвящены, соответственно, оценкам параметров статистических моделей и проверке гипотез. Они могут быть особенно полезны студентам при подготовке к экзамену по математической статистике.

Четвертая и пятая части (гл. 14–21) предназначаются, в первую очередь, лицам, желающим применить статистические методы для анализа экспериментальных данных.

Наконец, шестая часть (гл. 22–26) включает в себя ряд более специальных тем, обобщающих и дополняющих содержание предыдущих глав.

Собранный в книге материал неоднократно использовался на занятиях по математической статистике на механико-математическом факультете МГУ им. М. В. Ломоносова.

Автор будет считать свой труд небесполезным, если, перелистав книгу, читатель не потеряет к ней интереса, а захочет ознакомиться

Ей сна нет от французских книг, а мне от русских больно спится!

Фамусов в «Горе от ума» А. С. Грибоедова

Никогда не теряй из виду, что гораздо легче многих не удовлетворить, чем удовольствовать.

Козьма Прутков, «Мысли и афоризмы»

с теорией и приложениями статистики как по этому, так и по другим учебникам.

При работе над книгой образцом для автора была популярная серия книг для школьников Я. И. Перельмана. Хотелось, по возможности, использовать живую форму изложения и стиль, характерный для этой серии.

Я благодарен моим коллегам по лаборатории Математической статистики МГУ им. М. В. Ломоносова М. В. Козлову и Э. М. Кудлаеву за прочтение рукописи этой книги и полезные замечания.

М. Лагутин

К ЧИТАТЕЛЮ

Основа, подлинное содержание всякого познания доставляется именно наглядной концепцией мира, которая может быть добыта лишь нами самими и отнюдь не может быть как-либо преподана извне.

Артур Шопенгауэр, «Афоризмы житейской мудрости»

В книге Д. Пойа «Математическое открытие» (см. [62] в списке литературы) выделены *три принципа обучения*. Первым (и важнейшим) из них является

Стимулирование

Надо заинтересовать учащегося, убедить в полезности изучения предмета. Для успешности учебы необходимо четкое представление о том, зачем нужна сообщаемая информация.

Приведем мнение по этому вопросу известного героя детективного жанра (ведь восстановление по частностям общей картины есть также и задача математической статистики).

Студент — это не гусь, которого надо нафаршировать, а факел, который нужно зажечь.

> «Мне представляется, что человеческий мозг похож на маленький пустой чердак, который вы можете обставить, как хотите. Дурак натащит туда всякой рухляди, какая попадется под руку, и полезные, нужные вещи уже некуда будет всунуть, или в лучшем случае до них среди всей этой завали и не докопаешься. А человек толковый тщательно отбирает то, что он поместит в свой мозговой чердак. Он возьмет лишь инструменты, которые понадобятся ему для работы, но зато их будет множество, и все он разложит в образцовом порядке. Напрасно люди думают, что у этой маленькой комнатки эластичные стены и их можно растягивать сколько угодно. Уверяю вас, придет время, когда, приобретая новое, вы будете забывать что-то из прежнего. Поэтому страшно важно, чтобы ненужные сведения не вытесняли собой нужных.»
>
> *А. Конан Дойл, «Этюд в багровых тонах»*

Математическая статистика — один из наиболее часто используемых в приложениях разделов математики. На результаты практически любого научного эксперимента влияют неучтенные в модели факторы, накладывается случайный шум. Методы математической статистики, как правило, позволяют наиболее полно и надежно извлекать полезную информацию из зашумленных данных. В книгу включены многочисленные примеры применения статистических методов для решения практических задач.

Чтобы побудить читателя глубже изучить теорию вероятностей, на языке которой формулируются статистические теоремы, многие главы завершаются вероятностным парадоксом или занимательным экспериментом.

То, что вы были вынуждены открыть сами, оставляет в вашем уме дорожку, которой вы можете снова воспользоваться, когда в этом возникнет необходимость.

Г. Лихтенберг, «Aphorismen», Berlin, 1902–1906

Следующим принципом обучения является

Активность

По-настоящему разобраться в некоторой теории можно лишь самостоятельно решая задачи из данной области. Пассивного чтения даже хорошего учебника, увы, недостаточно для подлинного овладения предметом.

Каждая глава этой книги (за исключением дополнительных глав 22–26) содержит задачи (с решениями). Они обычно упорядочены по сложности, самые трудные отмечены звездочкой. Автор надеется, что читатель попробует решить некоторые из заинтересовавших его задач или, хотя бы, разберет решения, так как в них содержится значительная часть материала. Кроме того, по ходу изложения встречаются контрольные вопросы, ответы на которые приведены в конце соответствующей главы.

Возможность активного усвоения материала во многом определяется стилем его изложения.

При изложении математического рассуждения мастерство заключается в умении дать образованному читателю возможность сразу, не заботясь о деталях, схватить основную идею; последовательные дозы должны быть такими, чтобы их можно было глотать «с ходу»; в случае неудачи или если бы читатель захотел что-либо проверить, перед ним должна стоять четко ограниченная маленькая задача (например, проверить тождество; две пропущенные тривиальности могут в совокупности образовать непреодолимое препятствие).

Дж. Литлвуд, «Математическая смесь»

Наконец, третий принцип — это соблюдение последовательности фаз обучения

Исследование → формализация → усвоение

Важно начинать новую тему с содержательных примеров, чтобы можно было «потрогать руками», прочувствовать ситуацию. Можно попробовать придумать какой-нибудь способ решения проблемы лишь на основе здравого смысла. Если он на самом деле окажется бесполезным, то это лишь подтвердит важность теории, позволяющей получить приемлемое решение.

Всякое человеческое познание начинается с созерцаний, переходит от них к понятиям и заканчивается идеями.

И. Кант, «Критика чистого разума»

Абстрактные определения становятся по-настоящему понятны лишь тогда, когда они используются при решении конкретных задач в различных моделях. В книге «Теория катастроф» В. И. Арнольд пишет:

> «Абстрактные определения возникают при попытках обобщить «наивные» понятия, сохраняя их основные свойства. Теперь, когда мы знаем, что эти попытки не приводят к реальному расширению круга объектов (для многообразий это установил Уитни, для групп — Кэли, для алгоритмов — Черч), не лучше ли в преподавании вернуться к «наивным» определениям? (...) Пуанкаре подробно обсуждает методические преимущества наивных определений окружности и дроби в «Науке и методе»: невозможно усвоить правило сложения дробей, не разрезая, хотя бы мысленно, яблоко или пирог.»

При написании этой книги автор старался следовать указанным принципам обучения. Вероятно, какие-то методические приемы окажутся полезными преподавателям статистики, хотя, безусловно справедливо утверждал Козьма Прутков, что

У всякого портного свой взгляд на искусство!

Часть I

ВЕРОЯТНОСТЬ И СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Глава 1

ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ВЕЛИЧИН

В основе математической статистики лежит теория вероятностей. Аксиоматика теории вероятностей была разработана А. Н. Колмогоровым (опубликована в 1933 г.). Читателю, возможно, известны такие основные понятия этой теории, как независимость событий или математическое ожидание случайной величины. Тем не менее, будет полезно напомнить самое необходимое для дальнейшего изложения (см. также приложение П1*) и учебники [19], [39], [90] в списке литературы).

Вероятность — это важнейшее понятие в современной науке особенно потому, что никто совершенно не представляет, что оно означает.

Бертран Рассел, из лекции, 1929 г.

Читал ли что-нибудь? Хоть мелочь?

Репетилов в «Горе от ума» А. С. Грибоедова

Сперва аз да буки, а там и науки.

§ 1. ФУНКЦИИ РАСПРЕДЕЛЕНИЯ И ПЛОТНОСТИ

Пример 1. Измерим время ξ от первого включения до перегорания электрической лампочки.

Пример 2. Подбросим монетку. Если она упадет гербом вверх, будем считать, что $\xi = 1$, иначе положим $\xi = 0$.

Обобщая эти примеры, представим, что проводится эксперимент, результат которого (действительное число ξ) зависит от случая. Как охарактеризовать *случайную величину* ξ, дать вероятностный закон ее поведения?

Допустим, что возможно повторить эксперимент несколько раз. Обозначим через $\xi_1, \ldots, \xi_n$ полученные при этом значения. Тогда для произвольной точки x на прямой можно подсчитать ν_n — количество значений, попавших левее x (рис. 1).

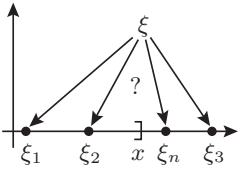

Рис. 1

Предположим, что существует некоторое число, к которому будет приближаться частота ν_n/n при неограниченном увеличении n. Естественно рассматривать это число как *вероятность того, что* ξ *не больше, чем* x. Обозначим эту вероятность через $\mathbf{P}(\xi \leqslant x)$. (Формальные определения понятий вероятности и случайной величины приведены в П1.)

P: Probabilitas (*лат.*) — вероятность.

Пример 3. На рис. 2 показан график частоты появлений буквы «а» в стихотворении М. Ю. Лермонтова «Бородино». Размах

*) П1 обозначает ссылку на раздел 1 приложения.

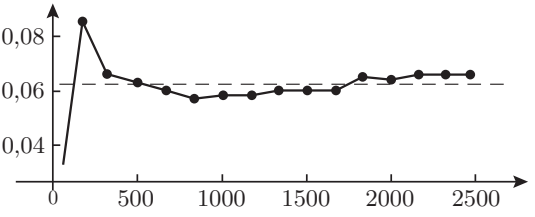

Рис. 2

колебаний частоты быстро уменьшается, она стабилизируется на уровне чуть большем, чем 0,06. В таблице приведены вероятности, с которыми встречаются в большом по объему тексте буквы русского алфавита, включая «пробел» между словами (данные взяты из [92, с. 238]). Отметим, что итоговая частота появлений буквы «а» в стихотворении «Бородино», равная $162/2461 \approx 0{,}066$, лишь незначительно отличается от соответствующей вероятности 0,062.

—	о	е, ё	а	и	т	н	с
0,175	0,090	0,072	0,062	0,062	0,053	0,053	0,045
р	в	л	к	м	д	п	у
0,040	0,038	0,035	0,028	0,026	0,025	0,023	0,021
я	ы	з	ь, ъ	б	г	ч	й
0,018	0,016	0,016	0,014	0,014	0,013	0,012	0,010
х	ж	ю	ш	ц	щ	э	ф
0,009	0,007	0,006	0,006	0,004	0,003	0,003	0,002

Зафиксируем n и рассмотрим поведение частоты ν_n/n при изменении «границы» x (см. рис. 1). При сдвиге точки x вправо, количество значений $\xi_1, \ldots, \xi_n$, оказавшихся левее x, будет увеличиваться. Поэтому вероятность $\mathbf{P}(\xi \leqslant x)$ (как предел частоты) будет неубывающей функцией от x, которая стремится к 1 при $x \to +\infty$ и стремится к 0 при $x \to -\infty$.

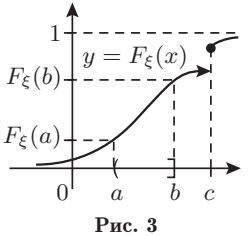

Рис. 3

Определение. Функция $F_\xi(x) = \mathbf{P}(\xi \leqslant x)$ называется *функцией распределения* случайной величины ξ.

Зная $F_\xi(x)$, можно найти вероятность попадания ξ в любой промежуток $(a,b]$ на прямой (рис. 3):

$$\mathbf{P}(a < \xi \leqslant b) = \mathbf{P}(\xi \leqslant b) - \mathbf{P}(\xi \leqslant a) = F_\xi(b) - F_\xi(a).$$

Вопрос 1.
Как это доказать формально, используя свойство непрерывности из П1?

Если функция распределения $F_\xi(x)$ имеет разрыв в точке c, то величина скачка $F_\xi(c) - F_\xi(c-)$ равна

$$\mathbf{P}(\xi = c) = \mathbf{P}(\xi \leqslant c) - \mathbf{P}(\xi < c).$$

Случайные величины мы будем задавать с помощью функций распределения.

Определение. Случайная величина η *равномерно распределена на отрезке* $[0,1]$, если

$$F_\eta(x) = \begin{cases} 0 & \text{при } x \leqslant 0, \\ x & \text{при } 0 < x < 1, \\ 1 & \text{при } x \geqslant 1. \end{cases}$$

Такое распределение соответствует *выбору точки наудачу* из отрезка $[0,1]$, поскольку для любых $0 \leqslant a < b \leqslant 1$ вероятность попадания значения η в отрезок $[a,b]$ равна его длине $b-a$ (рис. 4).

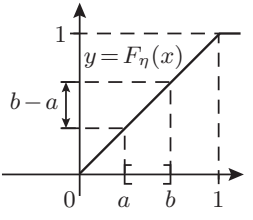

Рис. 4

Определение. Случайная величина τ называется *показательной с параметром* $\lambda > 0$, если

$$F_\tau(x) = \begin{cases} 0 & \text{при } x \leqslant 0, \\ 1 - e^{-\lambda x} & \text{при } x > 0. \end{cases}$$

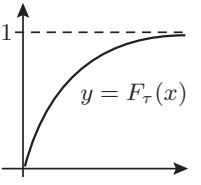

Рис. 5

График функции распределения $F_\tau(x)$ приведен на рис. 5.

Показательное распределение можно использовать для описания времени эксперимента из примера 1.

Вопрос 2.
Чему равна $\mathbf{P}(\tau > 3/\lambda)$ точно и приближенно?

Определение. Если существует такая функция $p_\xi(x) \geqslant 0$, что для произвольных $a < b$

$$\mathbf{P}(a \leqslant \xi \leqslant b) = \int_a^b p_\xi(x)\,dx,$$

то говорят, что случайная величина ξ (или ее распределение вероятностей) имеет *плотность* $p_\xi(x)$ (рис. 6).

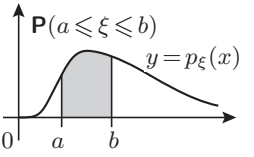

Рис. 6

Когда плотность существует, ее можно найти дифференцированием функции распределения:

$$p_\xi(x) = \frac{d}{dx}\,F_\xi(x) = \lim_{\Delta x \to 0} \frac{F_\xi(x+\Delta x) - F_\xi(x)}{\Delta x}\ .$$

Таким образом, плотностью равномерной величины η является функция $I_{[0,1]}$ (здесь и далее I_A обозначает *индикатор множества* A: $I_A(x) = 1$ при $x \in A$, $I_A(x) = 0$ при $x \notin A$), а плотностью показательной величины τ служит $p_\tau(x) = \lambda e^{-\lambda x} I_{[0,+\infty)}$ (рис. 7).

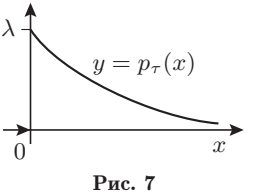

Рис. 7

Не у всякой случайной величины есть плотность. Например, ее нет у *дискретных* (принимающих конечное или счетное*) число значений) величин. Такова определяемая ниже бернуллиевская случайная величина.

Я. Бернулли (1654–1705), швейцарский математик.

*) Множество называют *счетным*, если его элементы можно перенумеровать натуральными числами.

Определение. Случайная величина ζ имеет *распределение Бернулли с вероятностью «успеха»* p $(0 \leqslant p \leqslant 1)$, если она принимает значения 0 и 1 с такими вероятностями: $\mathbf{P}(\zeta = 0) = 1 - p$ и $\mathbf{P}(\zeta = 1) = p$.

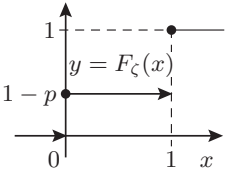

Рис. 8

График функции распределения $F_\zeta(x)$ бернуллиевской случайной величины ζ приведен на рис. 8. Распределение Бернулли при $p = 1/2$ годится как вероятностная модель эксперимента из примера 2. Значение $p \neq 1/2$ отвечает случаю несимметричной монеты.

Вопрос 3.
Как выглядит график функции распределения дискретной случайной величины ξ, принимающей значения $x_1 < x_2 < \ldots$ с соответствующими вероятностями $p_1, p_2, \ldots$?

§ 2. МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ И ДИСПЕРСИЯ

Не всегда требуется полная информация о случайной величине ξ, выражающаяся в ее функции распределения $F_\xi(x)$. Иногда достаточно знать, где располагается область «типичных» значений ξ. Одной из важных характеристик «центра» этой области является математическое ожидание.

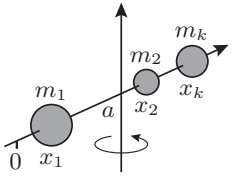

Рис. 9

Проблема. На тонком стержне (числовой прямой) в точках с координатами x_k находятся массы m_k (рис. 9). Где следует выбрать точку a крепления стержня к вертикальной оси, чтобы минимизировать *момент инерции* относительно нее $I_a = \sum (x_k - a)^2 m_k$?

Оказывается, точку крепления стержня надо поместить в *центр масс* $c = \sum x_k m_k / \sum m_k$ (см. задачу 1). Вероятностными аналогами центра масс c и момента инерции относительно него I_c служат математическое ожидание и дисперсия.

Определение. Для дискретной случайной величины ξ, принимающей значения $x_1, x_2, \ldots$ с соответствующими вероятностями $p_1, p_2, \ldots$, математическим ожиданием называется число

$$\mathbf{M}\xi = \sum_k x_k p_k. \tag{1}$$

Например, для бернуллиевской случайной величины ζ имеем

$$\mathbf{M}\zeta = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Определение. Когда у случайной величины ξ есть плотность $p_\xi(x)$, ее математическое ожидание вычисляется по формуле

$$\mathbf{M}\xi = \int_{-\infty}^{+\infty} x\, p_\xi(x)\, dx. \tag{2}$$

Для показательной случайной величины τ нетрудно подсчитать, интегрируя по частям, что

$$\mathbf{M}\tau = \int_0^\infty x\, \lambda e^{-\lambda x} dx = \frac{1}{\lambda} \int_0^\infty y\, e^{-y} dy = \frac{1}{\lambda} \left[0 + \int_0^\infty e^{-y} dy \right] = \frac{1}{\lambda}\,.$$

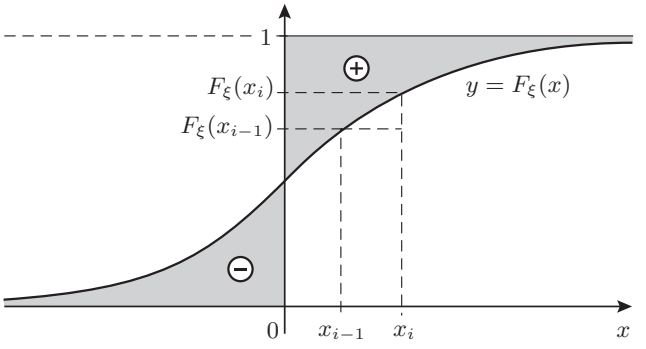

Рис. 10

Оба приведенных выше определения являются частными случаями следующего определения математического ожидания как интеграла Стильтьеса.

Определение. Для случайной величины ξ с функцией распределения $F_\xi(x)$ математическим ожиданием называется

$$\mathsf{M}\xi = \int\limits_{-\infty}^{+\infty} x\, F_\xi(dx) = \lim_{D\to 0} \sum_i x_i \,[F_\xi(x_i) - F_\xi(x_{i-1})],$$

где $D = \max |x_i - x_{i-1}|$ — *диаметр разбиения.*

Общее определение $\mathsf{M}\xi$ как интеграла Лебега приведено в приложении П2.

Рисунок 10 иллюстрирует *геометрическое представление* математического ожидания как разности площадей закрашенных областей со знаком «+» и знаком «−». Действительно, интегральная сумма в определении $\mathsf{M}\xi$ совпадает с суммой площадей (с учетом знака x_i) прямоугольников с шириной x_i и высотой $F_\xi(x_i) - F_\xi(x_{i-1})$. При измельчении разбиения она приближается к площади (с учетом знака) закрашенной области.

Геометрическое представление дает другой способ подсчета математического ожидания $\mathsf{M}\tau$ показательной случайной величины (см. рис. 5):

$$\mathsf{M}\tau = \int\limits_0^\infty \mathsf{P}(\tau > x)\, dx = \int\limits_0^\infty [1 - F_\tau(x)]\, dx = \int\limits_0^\infty e^{-\lambda x}\, dx = \frac{1}{\lambda}\,.$$

Для случайной величины ξ, принимающей только целые неотрицательные значения: $\mathsf{P}(\xi = k) = p_k$, $k \geqslant 0$, геометрическое представление величины $\mathsf{M}\xi$ (рис. 11) объясняет следующую формулу:

$$\mathsf{M}\xi = \sum_{k=0}^{\infty} \mathsf{P}(\xi > k). \qquad (3)$$

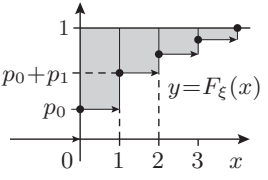

Рис. 11

Замечание. Математическое ожидание определено *не для всякой* случайной величины. Возможна ситуация, когда на рис. 10 и площадь области со знаком «+», и площадь области со знаком «−»

О. Коши (1798–1857), французский математик.

равны ∞. В этом случае возникает неопределенность вида $\infty-\infty$. Например, для *закона Коши* с плотностью $p_\xi(x)=1/[\pi\,(1+x^2)]$ каждая из площадей есть

$$\int\limits_0^\infty x\,p_\xi(x)\,dx=\frac{1}{\pi}\int\limits_0^\infty\frac{x}{1+x^2}\,dx=\frac{1}{2\pi}\int\limits_0^\infty\frac{dy}{1+y}=\frac{1}{2\pi}\,\ln(1+y)\Big|_0^\infty=\infty.$$

Следовательно, $\mathbf{M}\xi$ не существует, несмотря на то, что 0 — центр распределения (плотность $p_\xi(x)$ симметрична относительно 0).

Обсудим теперь понятие **дисперсии** случайной величины.

Как правило, помимо $\mathbf{M}\xi$ бывает важно знать величину типичного «разброса» значений ξ вокруг среднего. Мерой этого «разброса» может служить *стандартное отклонение* $\sqrt{\mathbf{D}\xi}$ (рис. 12), где *дисперсия* $\mathbf{D}\xi$ определяется формулой

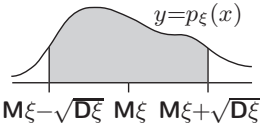

Рис. 12

$$\mathbf{D}\xi=\mathbf{M}\,(\xi-\mathbf{M}\xi)^2\,,$$

т. е. $\mathbf{D}\xi$ — это среднее квадрата отклонения ξ от $\mathbf{M}\xi$.

Для вычисления дисперсии полезно равенство

$$\mathbf{D}\xi=\mathbf{M}\xi^2-(\mathbf{M}\xi)^2\,. \tag{4}$$

Вопрос 4.
Как получить формулу (4) с помощью свойств математического ожидания из приложения П2?

Для примера вычислим дисперсию бернуллиевской случайной величины ζ. Прежде всего, заметим, что ζ^2 и ζ одинаково распределены. Поэтому $\mathbf{M}\zeta^2=\mathbf{M}\zeta=p$ и $\mathbf{D}\zeta=p-p^2=p(1-p)$.

Вы давиче его мне исчисляли свойства, но многие забыли? — Да?

Чацкий в «Горе от ума» А. С. Грибоедова [В слове «давеча» сохранена авторская орфография.]

§ 3. НЕЗАВИСИМОСТЬ СЛУЧАЙНЫХ ВЕЛИЧИН

Обычно вероятностную модель необходимо построить не для одного эксперимента, а для серии опытов. В этом случае нередко можно предполагать отсутствие взаимного влияния разных опытов друг на друга, их независимость.

Определение. Случайные величины $\xi_1,\ldots,\xi_n$ называются *независимыми*, если для любых $a_i<b_i$ $(i=1,\ldots,n)$

$$\mathbf{P}(a_i<\xi_i\leqslant b_i,\,i=1,\ldots,n)=\prod_{i=1}^n\mathbf{P}(a_i<\xi_i\leqslant b_i).$$

В частности, если все $a_i=-\infty$, то для произвольных $x_1,\ldots,x_n$

$$\mathbf{P}(\xi_1\leqslant x_1,\ldots,\xi_n\leqslant x_n)=F_{\xi_1}(x_1)\cdot\ldots\cdot F_{\xi_n}(x_n). \tag{5}$$

Независимые равномерно распределенные на отрезке $[0,1]$ случайные величины $\eta_1,\ldots,\eta_n$ можно считать координатами случайного вектора, равномерно распределенного в n-мерном единичном кубе. Действительно, равенство

$$\mathbf{P}(a_i\leqslant\eta_i\leqslant b_i,\ i=1,\ldots,n)=(b_1-a_1)\cdot\ldots\cdot(b_n-a_n),$$

где $0 \leqslant a_i < b_i \leqslant 1$, означает, что вероятность попадания точки $(\eta_1, \ldots, \eta_n)$ в произвольный параллелепипед с параллельными осям координат ребрами и находящийся целиком внутри единичного куба равна его объему (рис. 13 при $n = 3$). На самом деле, параллелепипед можно заменить на любое множество A, для которого определено понятие n-мерного объема.

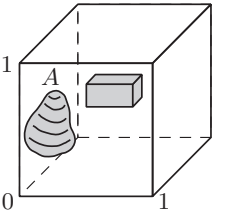

Рис. 13

Говорят, что бесконечная последовательность $\{\xi_i\}$ *образована независимыми случайными величинами*, если свойство независимости выполняется для любого конечного набора из них.

Определение. Последовательность независимых бернуллиевских случайных величин $\zeta_1, \zeta_2, \ldots$ с одинаковой вероятностью «успеха» p называют *испытаниями (или однородной схемой) Бернулли.*[*)]

В заключение параграфа приведем интуитивно понятное утверждение, которое часто применяется при доказательстве статистических теорем.

Лемма о независимости. Пусть $\xi_1, \ldots, \xi_{n+m}$ — независимые случайные величины; f и g — борелевские функции (см. приложение П2) на $\mathbb{R}^n$ и $\mathbb{R}^m$ соответственно. Тогда случайные величины $\eta_1 = f(\xi_1, \ldots, \xi_n)$ и $\eta_2 = g(\xi_{n+1}, \ldots, \xi_{n+m})$ независимы.

Доказательство этой леммы можно найти, например, в [48, с. 53].

§ 4. ПОИСК БОЛЬНЫХ

Применим элементарную теорию вероятностей к решению одной проблемы выявления больных (см. [82, с. 254]).

Во время второй мировой войны всех призывников в армию США подвергали медицинскому обследованию. Реакция Вассермана позволяет обнаруживать в крови больных сифилисом определенные антитела. Р. Дорфманом была предложена простая методика, на основе которой необходимое для выявления всех больных число проверок удалось уменьшить в 5 раз!

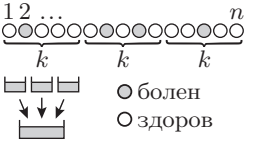

Рис. 14

Методика. Смешиваются пробы крови k человек и анализируется полученная смесь (рис. 14). Если антител нет, то этой одной проверки достаточно для k человек. В противном случае кровь каждого человека из этой группы нужно исследовать отдельно, и для k человек всего потребуется $k + 1$ раз провести анализ.

Вероятностная модель. Предположим, что вероятность обнаружения антител p одна и та же для всех n обследуемых, и результаты анализов для различных людей независимы, т. е. моделью является последовательность из n испытаний Бернулли с вероятностью «успеха» p.

Допустим для простоты, что n делится нацело на k. Тогда надо проверить n/k групп обследуемых. Пусть X_j — количество

[*)] Если у каждой случайной величины ζ_i своя вероятность «успеха» p_i, то схему называют *неоднородной*.

проверок, потребовавшихся в j-й группе, $j = 1, \ldots, n/k$. Тогда

$$X_j = \begin{cases} 1 & \text{с вероятностью } (1-p)^k \text{ (все } k \text{ человек здоровы)}, \\ k+1 & \text{с вероятностью } 1-(1-p)^k \text{ (есть больные)}. \end{cases}$$

Обозначим *общее число проверок* $X_1 + \ldots + X_{n/k}$ через Z. Задача заключается в том, как для заданного значения p[*)] определить размер группы $k_0 = k_0(p)$, минимизирующий $\mathbf{M}Z$.

Вопрос 5. Чем плох слишком большой размер группы?

Согласно формуле (1) находим

$$\mathbf{M}X_j = 1 \cdot (1-p)^k + (k+1) \cdot \left[1-(1-p)^k\right] = k+1-k\,(1-p)^k.$$

Отсюда по свойствам математического ожидания (П2) имеем

$$\mathbf{M}Z = \mathbf{M}X_1 + \ldots + \mathbf{M}X_{n/k} = \frac{n}{k}\ \mathbf{M}X_1 = n\left[1 + 1/k - (1-p)^k\right].$$

Вопрос 6. Какая ошибка допущена на рис. 15 в изображении графика функции $H(x)$ при малых p?

Положим $H(x) = 1 + 1/x - (1-p)^x$ при $x > 0$.

Для близких к нулю значений p минимум функции $H(x)$ достигается в точке x_0, где x_0 — наименьший из корней уравнения $H'(x) = 0$, т. е. уравнения

$$1/x^2 + (1-p)^x \ln(1-p) = 0. \tag{6}$$

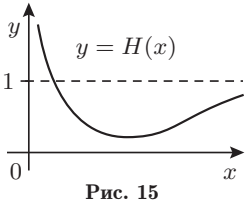

Рис. 15

Его нельзя разрешить явно относительно x. Поэтому, используя формулу $(1-p)^x \approx 1 - px$ при малых p, заменим $H(x)$ на функцию $\widetilde{H}(x) = 1 + 1/x - 1 + px = 1/x + px$, имеющую точку минимума $\widetilde{x}_0 = 1/\sqrt{p}$, причем $\widetilde{H}(\widetilde{x}_0) = 2\sqrt{p}$. Для $p = 0{,}01$ получаем $\widetilde{x}_0 = 10$ и $\widetilde{H}(\widetilde{x}_0) = 1/5$, т. е. $\mathbf{M}Z \approx n/5$.[**)]

Не пропускайте их, они еще не раз пригодятся в дальнейшем!

ЗАДАЧИ

1. Докажите, используя свойства математического ожидания (П2), что функция $f(a) = \mathbf{M}(\xi - a)^2$ при $a = \mathbf{M}\xi$ имеет минимум, равный $\mathbf{D}\xi$.

Я занимался до сих пор решением ряда задач, ибо при изучении наук примеры полезнее правил.

И. Ньютон, «Всеобщая арифметика»

2. Случайные величины $\eta_1, \ldots, \eta_n$ независимы и равномерно распределены на отрезке $[0, 1]$. Вычислите $\mathbf{M}\overline{\eta}$ и $\mathbf{D}\overline{\eta}$ среднего арифметического $\overline{\eta} = \dfrac{1}{n}\,(\eta_1 + \ldots + \eta_n)$.

3. Для случайных величин из задачи 2 найдите функцию распределения $F_{\eta_{(n)}}(x)$, $\mathbf{M}\eta_{(n)}$ и $\mathbf{D}\eta_{(n)}$, где $\eta_{(n)} = \max\{\eta_1, \ldots, \eta_n\}$.

4. Обозначим через ν число «неудач» до появления первого «успеха» в схеме Бернулли с параметром p. Вычислите $\mathbf{M}\nu$.
Указание. Примените формулу (3).

5. Рассмотрим следующую стратегию поиска больных. Все обследуемые разбиваются на пары. Если объединенная проба крови не содержит антител, то оба здоровы. В противном случае исследуется кровь первого из них. Если этот человек здоров, то другой должен быть болен, и в таком случае достаточно двух

[*)] Это значение можно оценить с помощью частоты выявления заболевания в предыдущих обследованиях.

[**)] Асимптотика x_0 и $H(x_0)$ при $p \to 0$ исследуется в задаче 6.

тестов. Если же первый оказался больным, то кровь второго также должна быть подвергнута анализу, и поэтому потребуется три теста. Выясните, при каких значениях вероятности p обнаружения заболевания у отдельного обследуемого данная стратегия будет в среднем экономичнее индивидуальной проверки.

6* Пусть $x_0 = x_0(p)$ — наименьший из корней уравнения (6). Докажите, что $x_0 \sim 1/\sqrt{p}$ и $H(x_0) \sim 2\sqrt{p}$ при $p \to 0$.[*)]

РЕШЕНИЯ ЗАДАЧ

1. С учетом свойств математического ожидания (см. приложение П2) и формулы (4) находим, что функция

$$f(a) = \mathbf{M}\left[\xi^2 - 2a\xi + a^2\right] = \mathbf{M}\xi^2 - 2a\mathbf{M}\xi + a^2 = (a - \mathbf{M}\xi)^2 + \mathbf{D}\xi$$

есть квадратный трехчлен с минимумом в точке $a = \mathbf{M}\xi$.

Растолковать прошу.

Репетилов в «Горе от ума» А. С. Грибоедова

2. Согласно формуле (2) $\mathbf{M}\eta_1 = \int\limits_0^1 x\,dx = 1/2$. (Это можно понять и без вычислений: плотность $p_{\eta_1}(x) = I_{[0,\,1]}$ симметрична относительно прямой $x = 1/2$.)

Далее, в силу следствия из П2 имеем $\mathbf{M}\eta_1^2 = \int\limits_0^1 x^2 dx = 1/3$. Применяя формулу (4), получаем, что $\mathbf{D}\eta_1 = 1/3 - 1/4 = 1/12$.

Наконец, согласно свойствам математического ожидания и дисперсии из приложения П2 запишем:

$$\mathbf{M}\overline{\eta} = \frac{1}{n}\left(\mathbf{M}\eta_1 + \ldots + \mathbf{M}\eta_n\right) = \mathbf{M}\eta_1 = \frac{1}{2},$$

$$\mathbf{D}\overline{\eta} = \frac{1}{n^2}\left(\mathbf{D}\eta_1 + \ldots + \mathbf{D}\eta_n\right) = \frac{1}{n}\,\mathbf{D}\eta_1 = \frac{1}{12\,n}$$

(во второй строке использована независимость случайных величин $\eta_1, \ldots, \eta_n$).

Обратим внимание на то, что случайные величины η_1 и $\overline{\eta}$ имеют одинаковое математическое ожидание, но дисперсия у $\overline{\eta}$ в n раз меньше. Эти соотношения, очевидно, выполняются и для произвольных независимых одинаково распределенных случайных величин $\varepsilon_1, \ldots, \varepsilon_n$ с конечной дисперсией. Такая модель используется для описания ошибок измерения.

Семь раз отмерь, а один — отрежь.

3. Максимум из случайных величин $\eta_1, \ldots, \eta_n$ не превосходит x тогда и только тогда, когда все η_i не больше, чем x (рис. 16), поэтому

$$F_{\eta_{(n)}}(x) = \mathbf{P}\left(\eta_{(n)} \leqslant x\right) = \mathbf{P}\left(\eta_1 \leqslant x, \ldots, \eta_n \leqslant x\right).$$

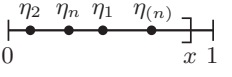

Рис. 16

[*)] Здесь $f(p) \sim g(p)$ означает, что $f(p)/g(p) \to 1$.

В силу независимости случайных величин η_i из формулы (5) для $x \in [0,1]$ выводим, что

$$F_{\eta_{(n)}}(x) = \mathbf{P}(\eta_1 \leqslant x) \cdot \ldots \cdot \mathbf{P}(\eta_n \leqslant x) = [\mathbf{P}(\eta_1 \leqslant x)]^n = x^n.$$

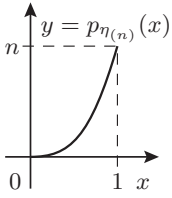

Рис. 17

График соответствующей плотности

$$p_{\eta_{(n)}}(x) = dF_{\eta_{(n)}}(x)/dx = nx^{n-1} I_{[0,1]}$$

изображен на рис. 17 (для $n > 2$).

Применяя формулу (2), вычисляем

$$\mathbf{M}\eta_{(n)} = \int_0^1 x\, nx^{n-1} dx = n \int_0^1 x^n dx = \frac{n}{n+1}.$$

Замечание. Интуитивно ясно, что длины отрезков, на которые делят $[0,1]$ взятые наудачу n точек, распределены одинаково (см. задачу 7 из гл. 10). Поэтому самая правая из точек будет находиться в среднем на расстоянии $1/(n+1)$ от 1. [Однако, *наименьший* из отрезков разбиения имеет длину порядка $1/n^2$ (задача 4 из гл. 4).]

Наконец, $\mathbf{M}\eta_{(n)}^2 = \int\limits_0^1 x^2\, nx^{n-1} dx = n/(n+2)$, откуда в силу соотношения (4) находим, что

$$\mathbf{D}\eta_{(n)} = n/(n+2) - [n/(n+1)]^2 = n/[(n+1)^2(n+2)].$$

Замечание. Дисперсия $\mathbf{D}\eta_{(n)}$ с ростом n убывает намного быстрее, чем дисперсия $\mathbf{D}\overline{\eta}$: порядок малости первой есть $1/n^2$, второй — $1/n$. Это связано с тем, что плотность $p_{\eta_1}(x) = I_{[0,1]}$ имеет разрыв в точке $x = 1$.

4. Вероятность p_k того, что до первого «успеха» в схеме Бернулли будет ровно k «неудач», в силу независимости испытаний равна $q^k p$, где $q = 1 - p$ (рис. 18). Это так называемое *геометрическое распределение.*[*)] Случайная величина ν дает пример дискретной случайной величины, имеющей счетное множество значений: $\mathbf{P}(\nu = k) = p_k$, $k \geqslant 0$. Суммируя геометрическую прогрессию, находим $\mathbf{P}(\nu > k) = p_{k+1} + p_{k+2} + \ldots = q^{k+1} p\,(1 + q + \ldots) = q^{k+1}$. Применяя формулу (3), получаем $\mathbf{M}\nu = q + q^2 + q^3 + \ldots = q/p$.

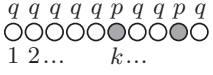

Рис. 18

5. Пусть Y_j — число проверок, потребовавшихся для j-й пары обследуемых ($j = 1,2,\ldots,n/2$), $q = 1 - p$. Тогда

$$Y_j = \begin{cases} 1 & \text{с вероятностью } q^2 \quad \text{(нет больных)}, \\ 2 & \text{с вероятностью } qp \quad \text{(первый здоров, второй болен)}, \\ 3 & \text{с вероятностью } (pq + p^2) = p \quad \text{(в противном случае)}. \end{cases}$$

[*)] Вероятности p_k образуют геометрическую прогрессию.