В. Босс

ЛЕКЦИИ *по* МАТЕМАТИКЕ

ТОМ 4

Вероятность, информация, статистика

МОСКВА

URSS

Оглавление

Предисловие к «Лекциям»

Самолеты позволяют летать, но добираться до аэропорта приходится самому.

Для нормального изучения любого математического предмета необходимы, по крайней мере, 4 ингредиента:

1) *живой учитель*;
2) *обыкновенный подробный учебник*;
3) *рядовой задачник*;
4) *учебник, освобожденный от рутины, но дающий общую картину, мотивы, связи, «что зачем»*.

До четвертого пункта у системы образования руки не доходили. Конечно, подобная задача иногда ставилась и решалась, но в большинстве случаев — при параллельном исполнении функций обыкновенного учебника. Акценты из-за перегрузки менялись, и намерения со второй-третьей главы начинали дрейфовать, не достигая результата. В виртуальном пространстве так бывает. Аналог объединения гантели с теннисной ракеткой перестает решать обе задачи, хотя это не сразу бросается в глаза.

«Лекции» ставят 4-й пункт своей главной целью. Сопутствующая идея — экономия слов и средств. Правда, на фоне деклараций о краткости и ясности изложения предполагаемое издание около 20 томов может показаться тяжеловесным, но это связано с обширностью математики, а не с перегрузкой деталями.

Необходимо сказать, на кого рассчитано. Ответ «на всех» выглядит наивно, но он в какой-то мере отражает суть дела. Обозримый вид, обнаженные конструкции доказательств, — такого

сорта книги удобно иметь под рукой. Не секрет, что специалисты самой высокой категории тратят массу сил и времени на освоение математических секторов, лежащих за рамками собственной специализации. Здесь же ко многим проблемам предлагается короткая дорога, позволяющая быстро освоить новые области и освежить старые. Для начинающих «короткие дороги» тем более полезны, поскольку облегчают движение любыми другими путями.

В вопросе «на кого рассчитано», — есть и другой аспект. На сильных или слабых? На средний вуз или физтех? Опять-таки выходит «на всех». Звучит странно, но речь не идет о регламентации кругозора. Простым языком, коротко и прозрачно описывается предмет. Из этого каждый извлечет свое и двинется дальше.

Наконец, последнее. В условиях информационного наводнения инструменты вчерашнего дня перестают работать. Не потому, что изучаемые дисциплины чересчур разрослись, а потому, что новых секторов жизни стало слишком много. И в этих условиях мало кто готов уделять много времени чему-то одному. Поэтому учить всему — надо как-то иначе. «Лекции» дают пример. Плохой ли, хороший — покажет время. Но в любом случае, это продукт нового поколения. Те же «колеса», тот же «руль», та же математическая суть, — но по-другому.

Предисловие к то́му

Без пассивной части словарного запаса — активная не работает.

Жизнь уходит на заделывание мелких трещин. Типографская краска — на уточнения. В теории вероятностей **(ТВ)** это особенно заметно из-за контраста простых выводов и сложных объяснений. Сложность, в свою очередь, проистекает из-за максималистских устремлений, согревающих профессионалов и убийственных для остальной части населения.

Учитывая, что профессионалы теорию вероятностей и так хорошо знают, нижеследующий текст ориентируется на умеренные аппетиты к строгости и детализации. Разумеется, обоснование того, что бóлее-менее «и так ясно», имеет свою цену. Но в **ТВ** на первом этапе гораздо важнее разобраться в том, что не ясно на самом элементарном уровне. Интуиция и здравый смысл настолько путаются в статистике и оценках вероятности, что многие тонкости вполне естественно отодвинуть на второй план.

Глава 1

Основы в задачах и парадоксах

1.1. Что такое вероятность

Карты, кости, тотализаторы, статистика, броуновское движение, отказы электроники, аварии, радиопомехи — все это вместе взятое создает ощущение смутного понимания случайности. Бытовое понятие вероятности оказывается в результате ничуть не менее основательным, чем бытовое понятие геометрической точки. Но стоит начать вдумываться, как явление ускользает. Теория вероятностей **(ТВ)** поэтому вглубь не идет, чтобы не угодить в ловушку.

> *Геометрия Евклида не определяет точек и прямых, шахматный кодекс — ферзя и пешек. Теория вероятностей не определяет, что такое вероятность элементарного события. Число от нуля до единицы. Первичное понятие, априори заданное. Вероятности сложных событий — другое дело. Этим, собственно, и занимается теория.*

Отправная точка у теории вероятностей очень проста. Рассматривается конечное или бесконечное множество [1)]

$$\Omega = \{\omega_1, \omega_2, \ldots\},$$

называемое ***пространством элементарных событий****, на котором задана функция* $p(\omega_i)$*, принимающая значения из* $[0, 1]$ *и удовлетворяющая условию нормировки* $\sum p(\omega_i) = 1$*. Значения* $p(\omega_i)$ *считаются вероятностями* ***элементарных событий*** ω_i*. Множества* $A \subset \Omega$ *называют событиями, и определяют их вероятности как*

$$\mathbf{P}(A) = \sum_{\omega_i \in A} p(\omega_i).$$

[1)] Пока счетное. Континуальные варианты Ω рассматриваются далее.

Вот и весь фундамент, упрощенно говоря. Исторический путь к нему был долгим и запутанным. Пройтись той же дорогой при изучении вероятностей, вообще говоря, необходимо. Пусть не в масштабе один к одному, но инкубационный период созревания понятий так или иначе должен быть преодолен.

Первым делом желательно привязать абстрактную модель к реальности. Для этого проще всего посмотреть, как примеры укладываются в общую схему.

Из колоды вытаскивается 7 карт. Какова вероятность, что среди них 3 короля и 2 дамы?

◀ Подтягивание задачи к общей схеме в данном случае совсем просто. Различные способы выбора 7 карт из 36 естественно считать равновероятными элементарными событиями, т. е.

$$p(\omega_i) = \frac{1}{C_{36}^7}, \quad \text{где} \quad C_n^k = \frac{n!}{k!(n-k)!}$$

— число сочетаний из n элементов по k элементов.

Число различных выборов, удовлетворяющих условиям задачи, равно $C_4^3 C_4^2 C_{28}^2$. Искомая вероятность есть $\dfrac{C_4^3 C_4^2 C_{28}^2}{C_{36}^7}$. ▶

В задачах, где элементарные события *равновероятны*, $\mathrm{P}(A)$ всегда равно числу вариантов, составляющих A, деленному на число всех вариантов:

$$\boxed{\mathrm{P}(A) = \frac{\textit{число благоприятных вариантов}}{\textit{число всех вариантов}}.}$$

На первый взгляд, суть дела тривиальна. Однако не все так просто, как поначалу кажется.

Парадокс Кардано[2]. При бросании двух шестигранных костей сумма выпавших чисел получается равной — как 9, так и 10 — в двух вариантах:

$$\text{сумма } 9 \Leftrightarrow (3,6)\ (4,5), \qquad \text{сумма } 10 \Leftrightarrow (4,6)\ (5,5).$$

Но вывод о равенстве вероятностей этих событий — ошибочен. Способов получения сумм 9 и 10 на самом деле больше, и их количество разное:

$$\text{сумма } 9 \Leftrightarrow (3,6)\ (6,3)\ (4,5)\ (5,4), \qquad \text{сумма } 10 \Leftrightarrow (4,6)\ (6,4)\ (5,5).$$

[2] Из «Книги об игре в кости», написанной Кардано в XXVI в., но изданной лишь в 1663 г.

Таким образом, из 36 возможных пар чисел 4 пары дают в сумме 9, и только 3 — 10. Вероятности, соответственно, равны 4/36 и 3/36, что подтверждает эксперимент[3].

На данном примере становится понятно, что в подборе пространства Ω элементарных событий имеется определенный произвол. Первый вариант — это 36 равновероятных *упорядоченных* пар (i, j). Второй вариант Ω — это *неупорядоченные* пары (21 пара), но тогда они не равновероятны, — и в этом аккуратно надо разобраться. Задача выглядит то простой, то сложной. Начинаешь присматриваться, и ум заходит за разум. Недаром Секей [22] отмечает, что в такого рода задачах ошибались в том числе великие (Лейбниц, Даламбер).

Путаницу в задаче создает независимость суммы от перестановки слагаемых. При последовательном выбрасывании костей — первая, потом вторая — проблемы не возникает. Но кости можно выбрасывать одновременно, они падают вместе, и первая от второй не отличается. Тогда различных вариантов имеется только 21 — и не вполне ясно, почему они не равновероятны[4].

Чтобы полностью развеять туман, полезно выделить подзадачу, в которой проблема сконцентрирована в максимально простом виде. *Какова вероятность при бросании двух костей получить в результате* (5,5) *и* (4,6)?

На примерах хорошо видна диалектика взаимоотношения события как смыслового явления, имеющего содержательное описание, и как множества из Ω, для выделения которого необходимо умение перечислить $\omega_i \in \Omega$, удовлетворяющие оговоренным в задаче условиям.

Реальность, окружающая абстрактную модель, включает ряд привходящих обстоятельств:

- возможность проведения опыта (эксперимента), исходом которого является наступление одного из элементарных событий[5] ω_i;

[3] При достаточно большом количестве бросаний двух костей — частоты, с которыми в сумме выпадают 9 и 10, стремятся к указанным вероятностям.

[4] Углубить непонимание можно, обратившись к парадоксу Гиббса в статистической физике. Смешение разнородных газов увеличивает энтропию. При естественном угле зрения не ясно, куда исчезает прирост энтропии, когда молекулы газов становятся одинаковы.

[5] Событие A наступает, если наступает $\omega_i \in A$.

- связь Ω с «субэлементарным уровнем» — с однократным бросанием монеты, например, тогда как элементарным событием может быть n-кратное бросание;
- возможность проведения серии опытов, в результате чего *частота наступления события A стремится к* $\mathbf{P}(A)$ *при увеличении длины серии*,

$$\frac{N(A)}{N} \to \mathbf{P}(A) \quad \text{при} \quad N \to \infty, \tag{1.1}$$

где N общее число опытов, а $N(A)$ число опытов, в которых наступило событие A.

Долгое время устойчивость частот была первична по отношению к понятию вероятности, и это в какой-то мере удовлетворяло спрос на понимание причин. Случившаяся затем метаморфоза изменила точку зрения на противоположную, но не ликвидировала выгод прежнего взгляда — ибо сходимость (1.1) превратилась в теорему и осталась в арсенале.

Вместе с тем с самого начала необходимо сказать о наличии логических трудностей — не в **ТВ**, но в непосредственной близости. Пусть речь идет о бросании монеты. Равенство вероятностей выпадения герба и решетки «вытекает», с одной стороны, из отсутствия оснований отдать предпочтение какой-либо альтернативе, с другой, — из наблюдения за длинными сериями бросаний.

Казалось бы, аргументов хватает. Тем не менее бросание монеты — хотя и сложная, но поддающаяся расчету механическая задача. По крайней мере, можно сконструировать высокоточный автомат, который почти всегда будет бросать монету гербом вверх. Почему же человек, действуя спонтанно, бросает «как надо»? Становится ясно, что источник случайности находится не в монете, а в человеке. Следующий вопрос ведет дальше, и причинно-следственная цепочка петляет по таким закоулкам Вселенной, что проблема, по большому счету, остается нерешенной.

1.2. Подводные рифы статистики

Теория оперирует вероятностями, практика — статистическими данными, т. е. исходами опытов, будь то бросание костей, количество аварий, смертей, выздоровлений, денег в казне и т. п. Умение делать выводы на базе статистики составляет оборотную сторону **ТВ**.

Не слишком утрируя действительность, допустим, что медики провели эксперимент по оценке влияния средства «чирикс» на заболевание «чикс». Как это всегда делается, контрольной группе давали плацебо. Гипотетические данные по Калуге и Рязани приведены в таблицах.

Калуга	*чирикс*	*плацебо*
помогло	10	1
безрезультатно	80	9

$$\Rightarrow \quad \frac{10}{10+80} > \frac{1}{1+9},$$

Рязань	*чирикс*	*плацебо*
помогло	10	89
безрезультатно	0	1

$$\Rightarrow \quad \frac{10}{0+10} > \frac{89}{1+89}.$$

Объединение результатов рождает химеру. В Калуге и Рязани чирикс эффективнее плацебо, в целом — наоборот.

Калуга + Рязань	*чирикс*	*плацебо*
помогло	20	90
безрезультатно	80	10

$$\Rightarrow \quad \frac{20}{20+80} < \frac{90}{10+90}.$$

На абстрактном уровне речь идет о следующем. Из

$$\frac{\xi_1}{\xi_1+\nu_1} > \frac{A_1}{A_1+B_1}, \qquad \frac{\xi_2}{\xi_2+\nu_2} > \frac{A_2}{A_2+B_2}$$

иногда делается поспешный вывод о справедливости неравенства

$$\frac{\xi_1+\xi_2}{\xi_1+\nu_1+\xi_2+\nu_2} > \frac{A_1+A_2}{A_1+B_1+A_2+B_2},$$

к чему нет никаких предпосылок.

Самое неприятное, что такого рода статистика — в облике экономических показателей и рейтингов — сваливается на нас со страниц вполне респектабельных газет.

1.3. Комбинаторика

Элементарная (но не обязательно простая) часть теории вероятностей в значительной мере опирается на комбинаторику.

Размещения. Число различных вариантов выбора (с учетом порядка) k предметов из n предметов $a_1, a_2, \ldots, a_n$ равно

$$\boxed{A_n^k = n(n-1)\ldots(n-k+1).}$$

◀ Есть n способов выбрать один предмет из n, т. е. $A_n^1 = n$. На каждый выбор первого предмета приходится $n-1$ возможностей выбора второго (из оставшихся $n-1$ предметов) — поэтому $A_n^2 = n(n-1)$. И так далее. ▶

Перестановки. Число всевозможных перестановок n предметов $a_1, \ldots, a_n$ равно «эн факториал»

$$\boxed{n! = 1 \cdot 2 \ldots n,}$$

что очевидно из $n! = A_n^n$.

По соображениям удобства принимается $0! = 1$.

Для оценки $n!$ при больших n удобна *формула Стирлинга*

$$\boxed{n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\theta_n/(12n)}, \qquad 0 < \theta_n < 1.}$$

Сочетания. Если k предметов из $a_1, \ldots, a_n$ выбираются без учета порядка (складываются в мешок), то число различных вариантов (число сочетаний из n по k) равно

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

◀ Всевозможные *размещения* получаются перестановками элементов в сочетаниях. Поэтому

$$A_n^k = C_n^k k!,$$

что дает формулу для C_n^k, с учетом того, что $A_n^k = n!/(n-k)!$ ▶

Перестановки с повторениями. Пусть имеется n предметов k типов

$$\underbrace{a_1 \ldots a_1}_{n_1}\underbrace{a_2 \ldots a_2}_{n_2}\ldots\underbrace{a_k \ldots a_k}_{n_k}, \qquad n_1 + \ldots + n_k = n.$$

Число различных перестановок этих предметов равно

$$\boxed{\mathbb{P}(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!}.}$$

◀ В любой перестановке рассматриваемой совокупности предметов, *ничего внешне не меняя*, можно n_1 элементов a_1 переставить между собой $n_1!$ способами,

n_2 элементов a_2 — $n_2!$ способами, ..., n_k элементов a_k — $n_k!$ способами. Поэтому $n_1!n_2!\ldots n_k!$ перестановок из $n!$ — неотличимы друг от друга, что приводит к указанной формуле. ▶

В слове «абракадабра» 5 букв «а», 2 — «б», 2 — «р», 1 — «к», 1 — «д». Из такого набора букв можно сделать

$$\mathbb{P}(5,2,2,1,1) = \frac{11!}{5!2!2!} = 83010$$

различных буквосочетаний.

Выбор из k типов. Имеется k типов предметов, каждый тип представлен бесконечным количеством экземпляров. Число различных способов выбора r предметов в данном случае

$$\boxed{U_k^r = k^r.}$$

Ситуация из перечисленных самая простая, но иногда почему-то ставит в тупик. Десять (типов) цифр, шестизначных чисел — миллион, 10^6.

Упражнения [6)]

- Сколько различных чисел можно получить перестановкой четырех цифр 1, 3, 5, 7? (4!).
- Сколько есть восьмизначных чисел, в записи которых участвуют только цифры 1, 3, 5, 7? (4^8).
- Сколько есть различных чисел, в записи которых участвуют две единицы и одна семерка? (3).
- При размещении n шаров по n ячейкам вероятность того, что все ячейки будут заняты, равна $n!/n^n$.
- При размещении k шаров (дней рождения) по 365 ячейкам (дням) вероятность того, что все шары попадут в разные ячейки, равна $A_{365}^k/365^k$.

1.4. Условная вероятность

Объединение и пересечение событий. *Объединением* или *суммой событий* A и B называют событие, состоящее в наступлении хотя бы одного из событий A, B и обозначаемое как $A \cup B$ или $A+B$. Первое обозначение прямо указывает, какое множество в Ω отвечает сумме событий.

[6)] Упражнения в «Лекциях» используются в основном как способ поместить в фокус внимания некоторые факты без обсуждения деталей.

Пересечением или *произведением событий* A и B называют событие, состоящее в совместном наступлении A, B и обозначаемое как $A \cap B$ или AB.

Очевидно,

$$\boxed{\mathbf{P}(A+B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(AB),} \qquad (1.2)$$

поскольку при суммировании ω_i по A и B элементарные события из пересечения AB считаются два раза, и один раз $\mathbf{P}(AB)$ приходится вычесть. Если события не пересекаются, то

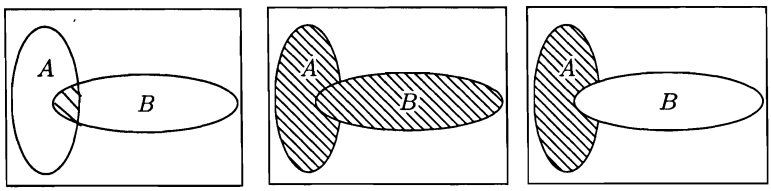
$$\mathbf{P}(A+B) = \mathbf{P}(A) + \mathbf{P}(B).$$

Формулы типа (1.2) становятся совершенно прозрачны при использовании рисунков объединения и пересечения множеств (рис. 1.1). Опробовать рецепт можно на проверке равенства

$$\mathbf{P}(A+B+C) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(AB) - \mathbf{P}(AC) - \mathbf{P}(BC) + \mathbf{P}(ABC),$$

а также в общем случае n событий $A_1, \ldots, A_n$:

$$\mathbf{P}\Big(\sum_k A_k\Big) = \sum_k \mathbf{P}(A_k) - \sum_{i,j} \mathbf{P}(A_i A_j) + \sum_{i,j,k} \mathbf{P}(A_i A_j A_k) - \ldots . \qquad (1.3)$$



AB $A+B$ $A-B$

Рис. 1.1

Параллели логических высказываний с операциями над множествами используются достаточно широко. Событию «не A» отвечает дополнение $\overline{A}$ множества A в Ω, а разность $A \setminus B$, или $A - B$, интерпретируется как наступление A, но не B. Наконец, *симметрическая разность*

$$A \triangle B = (A \cup B) \setminus (A \cap B)$$

обозначает событие, состоящее в наступлении одного из A, B, но не двух вместе.

Пустое множество $\varnothing$, считается, принадлежит Ω и символизирует невозможное событие. При этом $\mathbf{P}(\varnothing) = 0$.

С учетом нормировки $\mathbf{P}(\Omega) = 1$, очевидно, $\mathbf{P}(A) + \mathbf{P}(\overline{A}) = 1$.

Перечисленные действия над событиями в совокупности с формулами вычисления вероятностей позволяют решать многие задачи, не спускаясь на уровень рассмотрения пространства элементарных событий. Это экономит усилия, но иногда затрудняет ориентацию.

Условная вероятность. Вероятность $\mathbf{P}(B|A)$ наступления B при условии наступления в то же время события A — называют *условной*.

Из всех $\omega_i \in A$ входят в B лишь ω_i, принадлежащие пересечению AB. Они-то и определяют $\mathbf{P}(B|A)$. И если бы A было нормировано, то $\mathbf{P}(B|A)$ равнялось бы $\mathbf{P}(AB)$. Нормировка A корректирует результат очевидным образом:

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(AB)}{\mathbf{P}(A)}. \tag{1.4}$$

Перезапись (1.4) в форме

$$\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B|A) \tag{1.5}$$

называют *формулой умножения вероятностей*.

Задача. *Имеется три картонки. На одной — с обеих сторон нарисована буква* А, *на другой —* В. *На третьей картонке с одной стороны* А, *с другой —* В. *Одна из картонок выбирается наугад и кладется на стол. Предположим, на видимой стороне картонки оказывается буква* А. *Какова вероятность, что на другой стороне — тоже* А*?*

«Одна вторая», — ошибочно отвечает интуиция, и причина заблуждения далеко не очевидна. Дело в том, что картонка не только случайно выбирается, но и случайно укладывается на одну из сторон. Поэтому логика здесь такая. Всего имеется шесть нарисованных букв, из них — три буквы А, две на картонке АА и одна — на АВ. Букву А из АА вытащить в два раза более вероятно, чем из АВ. Получается, вероятность того, что на столе лежит картонка АА, равна 2/3.

Если кого-то смущают картонки, то это — для простоты и краткости. Реальные прикладные задачи описывать громоздко, а читать скучно. Но таких задач, где здравый смысл терпит фиаско, довольно много. И дело не в том, что ахиллесова пята интуиции приходится на вероятность. Слабое место интуиции в другом. Взаимодействие всего двух факторов ставит воображение в тупик. А комбинация многофакторности с наглядностью — в теории вероятностей такова, что все время искрит.

Формула Байеса. Разбиение Ω на *полную группу несовместимых*[7] *событий* $A_1, \ldots, A_n$ позволяет любое событие B записать в виде

$$B = BA_1 + \ldots + BA_n,$$

откуда $\mathbf{P}(B) = \mathbf{P}(BA_1) + \ldots + \mathbf{P}(BA_n)$, и в силу (1.5) — получается *формула полной вероятности*:

$$\boxed{\mathbf{P}(B) = \mathbf{P}(B|A_1)\mathbf{P}(A_1) + \ldots + \mathbf{P}(B|A_n)\mathbf{P}(A_n).} \tag{1.6}$$

Пусть $\mathbf{P}(A), \mathbf{P}(B) > 0$. Из

$$\mathbf{P}(AB) = \mathbf{P}(A|B)\mathbf{P}(B) = \mathbf{P}(B|A)\mathbf{P}(A)$$

вытекает

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\mathbf{P}(A)}{\mathbf{P}(B)},$$

что после учета (1.6) приводит к *формуле Байеса*

$$\mathbf{P}(A_j|B) = \frac{\mathbf{P}(B|A_j)\mathbf{P}(A_j)}{\sum\limits_k \mathbf{P}(B|A_k)\mathbf{P}(A_k)}, \tag{1.7}$$

сильно скомпрометированной безосновательными попытками ее применения.

Неутихающее колоброжение вокруг (1.7) всегда определяла соблазнительная интерпретация. Если A_j это гипотезы с априорными вероятностями $\mathbf{P}(A_j)$, то при наступлении события B в результате эксперимента — формула определяет апостериорные вероятности $\mathbf{P}(A_j|B)$. Звучит красиво, но априорные вероятности, как правило, не известны. А поскольку с идеей расставаться жалко, $\mathbf{P}(A_j)$ начинают трактовать как степень уверенности.

1.5. Случайные величины

Числовая функция[8] $X(\omega)$, заданная на Ω, представляет собой *случайную величину* (с. в.). Примером может служить функция, принимающая значения 1 или 0 при выпадении герба или решетки.

[7] Непересекающихся.

[8] О необходимых уточнениях см. главу 10.

Среднее значение $m_x = \mathbf{E}(X)$,

$$\boxed{\mathbf{E}(X) = \sum_{\omega \in \Omega} X(\omega)\mathbf{P}(\omega),}$$

называют *матожиданием*[9] $X(\omega)$.

Математическое ожидание *функции-индикатора* $\chi_A(\omega)$ множества A,

$$\chi_A(\omega) = \begin{cases} 1, & \text{если } \omega \in A; \\ 0, & \text{если } \omega \notin A, \end{cases}$$

равно, очевидно, вероятности $\mathbf{P}(A)$.

Матожидание представляет собой весьма важную характеристику случайной величины. Очевидно,

$$\mathbf{E}(\alpha X + \beta Y) = \alpha \mathbf{E}(X) + \beta \mathbf{E}(Y).$$

Еще можно отметить: $X(\omega) \geqslant 0 \Rightarrow \mathbf{E}(X) \geqslant 0$.

На вид все очень просто, но, как говорится, в тихом омуте черти водятся.

Парадокс транзитивности. Сравнивая случайные величины X и Y, будем говорить *«X больше Y по вероятности»*, — если

$$\mathbf{P}\{X > Y\} > \mathbf{P}\{X \leqslant Y\},$$

т. е. вероятность неравенства $X > Y$ больше $1/2$.

Пусть пространство элементарных событий Ω состоит из 6 точек, в которых с. в. X, Y, Z, W с равной вероятностью $1/6$ принимают значения согласно таблице[10]:

X	6	6	2	2	2	2
Y	5	5	5	1	1	1
Z	4	4	4	4	0	0
W	3	3	3	3	3	3

Очевидно, $X = 6$ с вероятностью $1/3 = 2/6$. В этом случае $X > Y$ независимо от значения Y. С вероятностью $2/3 = 4/6$ величина X равна 2. Тогда $X > Y$,

[9] Математическим ожиданием.

[10] Функция X, например, может быть реализована бросанием шестигранной кости, грани которой помечены цифрами {6 6 2 2 2 2}.

если $Y = 1$, что имеет вероятность $1/2 = 3/6$. Поэтому, с учетом формул умножения вероятностей и суммы непересекающихся событий, итоговая вероятность неравенства $X > Y$ равна

$$\frac{1}{3} + \frac{2}{3} \cdot \frac{1}{2} = \frac{2}{3}.$$

Аналогично подсчитывается, что $Y > Z$, $Z > W$, — с той же вероятностью $2/3$. Получается цепочка неравенств

$$X > Y > Z > W.$$

Возможность $W > X$ представляется в некотором роде дикой. Тем не менее $W > X$ с вероятностью $2/3$ (!).

Парадокс ожидания серии. Какая в случайной «01»-последовательности[11] комбинация, 00 или 01, появится раньше? Очевидно, равновероятно, поскольку после первого появления нуля на следующем шаге возникнет либо 0, либо 1, — с вероятностью $1/2$.

Напрашивается вывод, что среднее число шагов (среднее время ожидания) m_{00} и m_{01} до появления, соответственно, серий 00 либо 01 — тоже одинаково. Но это не так.

◀ Пусть m_0 обозначает среднее число шагов до появления комбинации 01 при условии, что первая цифра «01»-последовательности оказалась *нулем*, а m_1 — среднее число шагов до появления комбинации 01 при условии, что первая цифра «01»-последовательности оказалась *единицей*. Легко видеть, что

$$m_0 = 1 + \frac{1}{2} + \frac{1}{2}m_0, \qquad m_1 = 1 + \frac{1}{2}m_0 + \frac{1}{2}m_1,$$

откуда

$$m_0 = 3, \qquad m_1 = 5, \qquad m_{01} = \frac{m_0 + m_1}{2} = 4.$$

Если же m_0^*, m_1^* обозначают аналоги m_0, m_1 в ситуации, когда речь идет о появлении комбинации 00, то

$$m_0^* = 1 + \frac{1}{2} + \frac{1}{2}m_1^*, \qquad m_1^* = 1 + \frac{1}{2}m_0^* + \frac{1}{2}m_1^*.$$

В конечном итоге это дает $m_{00} = \dfrac{m_0^* + m_1^*}{2} = 6$. Писать опровержения можно во многие адреса[12]. ▶

Не так удивительно, но заслуживает упоминания, что из

«$U > V$ по вероятности»,

[11] Подразумевается равная вероятность появления нуля и единицы.

[12] Дополнительную информацию можно найти в [22], см. также *Li Shou-Yen R.* A martingale approach to the study of occurrence of sequence patterns in repeated experiments // Annals of Prob. 1980. **8**. P. 1171–1176.

вообще говоря, не следует $\mathbf{E}\{U\} > \mathbf{E}\{V\}$, но $U(\omega) > V(\omega)$, конечно, влечет за собой $\mathbf{E}\{U\} > \mathbf{E}\{V\}$.

Заслуживает упоминания также оборотная сторона медали. **ТВ**, как сильнодействующее средство, не только спасает от заблуждений, но и создает их.

Общепринято думать, например, что в лотерею играть неразумно, поскольку матожидание выигрыша меньше стоимости билета. В результате покупать лотерейные билеты приходится, оглядываясь по сторонам.

При этом подсознательно все понимают, что конечный денежный выигрыш может иметь бесконечную ценность. Покупка дома, переезд, лечение, образование. Да мало ли что еще меняет судьбу, и потому в деньгах не измеряется, хотя нуждается в той или иной стартовой сумме. Почему же за 30 копеек не купить шанс? Взвешивание здесь только вредит. Но авторитет иероглифов формул и таинственной терминологии создает гипнотизирующий мираж.

Другой пример — знаменитый *«Петербургский парадокс»*. Если герб при неоднократном бросании монеты выпадает в первый раз в n-й попытке, — участнику игры выплачивается 2^n рублей.

Математическое ожидание выигрыша,

$$2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + \ldots + 2^n \cdot \frac{1}{2^n} + \ldots = 1 + 1 + \ldots,$$

бесконечно. Поэтому, с точки зрения ТВ (как бы), за участие в игре денег можно заплатить сколько угодно — казино в любом случае проиграет.

Хороший пример на тему того, как респектабельная теория направляет ход мыслей не в то русло, тогда как реальная задача не стоит выеденного яйца. Казино проигрывает в среднем, но в данном случае это не дает разумных оснований судить об одноразовой игре. Средние значения продуктивно работают в других ситуациях, но не здесь.

Рассмотрим упрощенный аналог. Монета бросается один раз, и падает плашмя с вероятностью $1 - 2^{-100}$. Выигрыш при этом составляет 1 рубль. На ребро монета становится с вероятностью 2^{-100}, и тогда выигрыш равен 2^{300} рублей. Матожидание выигрыша $\sim 2^{200}$. Но, очевидно, больше рубля за участие в игре платить глупо. Потому что событие, имеющее вероятность 2^{-100} «никогда» не случается, и какая разница, сколько за него обещано. Использование матожидания оказывается просто не к месту.

В русле «Петербургского парадокса» было сломано немало копий при участии великих математиков. В рамках идеологии сходимости (глава 4) это довольно типичная ситуация, когда X_n сходится по вероятности к нулю, а матожидание X_n стремится к бесконечности.

1.6. Континуальные пространства

Пространство элементарных событий Ω часто имеет континуальную природу. Это может быть вещественная прямая или отрезок, R^n либо его подмножество. За кадром здесь находятся вполне естественные задачи. Вот два рядовых примера.

- *Стержень AB ломается в точках P и Q на три куска. Какова вероятность того, что из них можно сложить треугольник?*

◄ В случае $x = AP$, $y = PQ$ возможность сложить треугольник описывается неравенствами

$$\frac{l}{2} < x + y < l; \quad x, y < \frac{l}{2},$$

которым на рис. 1.2 удовлетворяют внутренние точки треугольника EFG. Если все точки $\{x, y\}$ равновероятны, то искомая вероятность

$$P = \frac{S_{\triangle EFG}}{S_{\triangle OCD}} = \frac{1}{4}. \quad ►$$

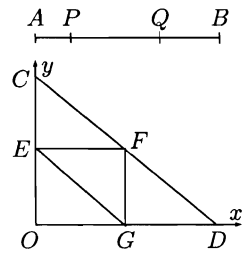


Рис. 1.2

- *Во время боя в течение часа в корабль попадает два снаряда. Для заделки одной пробоины требуется 15 минут. Если пробоина еще не заделана, а в корабль попадает второй снаряд, — корабль тонет. Какова вероятность потопить корабль?*

◄ Если времена попаданий снарядов t_1 и t_2 равномерно распределены по квадрату S размера 60 мин $\times$ 60 мин, то искомую вероятность дает отношение площади многоугольника $\{|t_1 - t_2| \leqslant 15\} \cap S$ к площади S. ►

Задачи подобного рода берут начало от классической *задачи Бюффона* об игле [13] и предполагают, как правило, равномерность распределения параметров в некоторой области Ω. Вероятность попадания точки в подмножество $A \subset \Omega$ считается при этом равной отношению площадей S_A к S_Ω, что выглядит достаточно логично. Безмятежное отношение к такой идеологии сохранялось до столкновения с *парадоксом Бертрана* [14].

В *задаче Бертрана* вычисляется вероятность того, что наугад взятая хорда заданной окружности больше стороны вписанного правильного треугольника.

[13] Игла длиной r бросается на плоскость, разграфленную параллельными прямыми, отстоящими друг от друга на расстоянии $a > r$. Какова вероятность того, что игла пересечет одну из параллелей? Ответ: $p = 2r/(a\pi)$.

[14] *Bertrand J.* Calcul des probabilites. P., 1889.

Бертран рассмотрел три варианта параметров, определяющих положение хорды:

- расстояние до центра и угол нормали хорды с осью x;
- угловые координаты точек пересечения хорды с окружностью;
- декартовы координаты середины хорды.

Во всех трех случаях вероятности оказались разными (1/2, 1/3, 1/4).

Казус пошел на пользу. Стало ясно, что внимательнее надо изучать ситуацию неравномерного распределения точек в Ω.

Допустим, точки в Ω распределены с плотностью $\mu(\omega)$, причем

$$\int_{\Omega} \mu(\omega)\, d\omega = 1.$$

Тогда вероятность события $\omega \in A$ определяется как

$$\mathbf{P}(A) = \int_{A} \mu(\omega)\, d\omega,$$

а если на Ω задана случайная величина $X(\omega)$, в том числе векторная $X(\omega) = \{X_1(\omega), \ldots, X_n(\omega)\}$, то матожидание равно

$$\mathbf{E}\,(X) = \int_{\Omega} X(\omega)\mu(\omega)\, d\omega. \tag{1.8}$$

Сделаем теперь важный шаг. Изменим точку отсчета. Первичное пространство элементарных событий сыграло свою роль, и при необходимости можно обойтись без него. Во многих ситуациях это дает определенные выгоды.

Урожайность капусты, например, случайная величина? По-видимому. Но на нее влияет столько факторов, что о наличии глубинного Ω мы можем только догадываться. Господь бросает кости по ту сторону, а мы тут наблюдаем результат — саму случайную величину. И даже в простейшем случае бросания монеты — пространство «герб — решетка» лишь агрегированная иллюзия. Реальное Ω надо искать на другом уровне, где об устройстве Вселенной известно немного больше.

Исключить исходное Ω из рассмотрения можно, переходя непосредственно на описание с. в. X с помощью *функции распределения*:

$$\boxed{F(x) = P(X < x).}$$

Разумеется,

$$\mathbf{P}(X < x) = \int\limits_{X<x} \mu(\omega)\, d\omega,$$

но это остается за кадром. Таким образом, случайные величины могут характеризоваться непосредственно в терминах функций распределения. *Отказ от рассмотрения пространства элементарных событий носит, разумеется, условный характер. На самом деле одно пространство заменяется другим.* Происходит нечто вроде агрегирования. Пространством Ω случайной величины X становится вещественная прямая или ее подмножество. Вне поля зрения остается более глубокий уровень, если таковой имеется.

Очевидно, функция $F(x)$ монотонно возрастает (не убывает) и

$$\lim_{x\to\infty} F(x) = 1, \qquad \lim_{x\to-\infty} F(x) = 0.$$

Вместо $F(x)$ часто используют *плотность распределения* $\rho(x)$, связанную с $F(x)$ условием:

$$F(x) = \int\limits_{-\infty}^{x} \rho(u)\, du. \tag{1.9}$$

Из (1.9) следует

$$F(x+\Delta x) - F(x) = \int\limits_{x}^{x+\Delta x} \rho(u)\, du = \rho(x)\Delta x + o(\Delta x),$$

откуда

$$\boxed{\rho(x) = F'(x).}$$

Понятно, что для дифференцируемости $F(x)$ нужны предположения, но мы на этом не останавливаемся. Более того, далее используются — в том числе — плотности, содержащие δ-функции [15], что позволяет единообразно охватить дискретно и непрерывно распределенные случайные величины.

[15] См. [5, т. 2].

Аналогом равновероятных элементарных событий служит ситуация равномерной плотности:

$$\rho(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b]; \\ 0, & x \notin [a, b]. \end{cases}$$

При этом говорят о *равномерном распределении* X на $[a, b]$.

Если X вектор, то в $F(x) = P(X < x)$ под $X < x$ подразумевается совокупность покомпонентных неравенств. Из

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} \rho(u_1, \ldots, u_n)\, du_1 \ldots du_n$$

вытекает

$$\rho(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \ldots x_n}.$$

Вместо $F(x)$ и $\rho(x)$ обычно пишут $F_X(x)$ и $\rho_X(x)$, помечая случайную величину X и аргумент x. Мы не только будем опускать индекс, но вместо X будем иногда писать x. Конечно, это не вполне корректно, но не более чем $\int^x f(x)\, dx$. Строгие обозначения — не всегда благо. Если из контекста ясно, о чем речь, обозначение тем лучше — чем проще.

В соответствии со сказанным [16)],

$$\mathbf{E}(X) = \int x\rho(x)\, dx.$$

При нежелании впадать в обсуждение деталей то же самое пишут в виде

$$\mathbf{E}(X) = \int x\, dF(x).$$

[16)] Бесконечные пределы при интегрировании иногда опускаются, — в результате $\int$ обозначает $\int_{-\infty}^{\infty}$. Еще лучше сказать, что $\int$ обозначает интегрирование по области определения функции, стоящей под интегралом.

Далее хотелось бы сказать, что линейность оператора $\mathbf{E}$,

$$\mathbf{E}(\alpha X + \beta Y) = \alpha \mathbf{E}(X) + \beta \mathbf{E}(Y), \tag{1.10}$$

очевидна. Но это не совсем так. В ситуации (1.8) подразумевалось, что случайные величины могут быть разные, но мера $\mu(\omega)$ на Ω одна и та же, — и тогда линейность действительно очевидна. В данном случае X, Y могли «прибыть» на $(-\infty, \infty)$ из разных Ω, к тому же предысторией уже никто не интересуется, X имеет свою плотность распределения, Y — свою. Точнее говоря, надо рассматривать даже совместную плотность их распределения $\rho(x, y)$. Тогда

$$\mathbf{E}(\alpha X + \beta Y) = \iint (\alpha x + \beta y)\rho(x, y)\, dx\, dy =$$

$$= \alpha \int x\rho(x)\, dx + \beta \int y\rho(y)\, dy = \alpha \mathbf{E}(X) + \beta \mathbf{E}(Y),$$

где, например, $\rho(x) = \int \rho(x, y)\, dy$.

Пояснение простого факта может показаться многословным, но здесь имеет смысл потратить какое-то время, если речь идет о попытке осмыслить скелетную основу **ТВ**. Исходное определение с. в. создает впечатление, что Ω задано, и время от времени в рассмотрение включаются разные случайные величины. Реальная картина обычно другая. Каждая с. в. приходит как бы со своей прямой $(-\infty, \infty)$, и Ω постепенно расширяется с $(-\infty, \infty)$ до $(-\infty, \infty) \times (-\infty, \infty)$ и т. д.

Добавление к линейности $\mathbf{E}$ трех аксиом типа $\mathbf{E}(1) = 1$ позволяет использовать матожидание как отправную точку — вместо вероятности. Соответствующий сценарий описан в книге Уиттла [25]. Разумеется, при упрощенном взгляде на предмет особой математической разницы нет, поскольку вероятность возникает тут же как матожидание функции-индикатора,

$$\mathbf{P}(A) = \mathbf{E}[\chi_A(X)].$$

Разница подходов начинает ощущаться на высоких этажах теории вероятностей, где круг дозволенных неприятностей достаточно широк.

Но помимо математической — есть разница психологическая. Для кого-то понятие среднего значения может быть предпочтительнее понятия вероятности.

Пример. Случайная величина, принимающая значения из ограниченного промежутка, всегда имеет матожидание. При распределении на бесконечном промежутке — не обязательно. Пусть с. в. X распределена по *закону Коши*, характеризуемому плотностью

$$\rho(x) = \frac{1}{\pi(1 + x^2)}.$$

Тогда

$$m_x = \int \frac{x\, dx}{\pi(1 + x^2)} = \infty.$$

При переходе к моментам более высокого порядка ситуация только ухудшается.

1.7. Независимость

События A и B называют *независимыми*, если $\mathbf{P}(B|A) = \mathbf{P}(B)$, т. е. формула умножения вероятностей (1.5) переходит в

$$\boxed{\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B).} \tag{1.11}$$

Из (1.11), в свою очередь, следует $\mathbf{P}(A|B) = \mathbf{P}(A)$.

Понятие независимости играет фундаментальную роль в теории вероятностей, но (1.11) не вполне отвечает интуитивному пониманию независимости, что имеет смысл сразу оговорить.

Парадокс Бернштейна. Бросают две монеты. Пусть выпадение первой монеты гербом обозначает событие A, второй — B. Наконец, C означает, что только одна монета выпала гербом.

Для симметричных монет все три события попарно независимы, поскольку

$$P(A) = P(B) = P(C) = \frac{1}{2}, \qquad P(AB) = P(AC) = P(BC) = \frac{1}{4}. \tag{1.12}$$

С независимостью A и B интуиция согласна, но не с независимостью A и C (или B и C). И у нее есть основания. Независимость (1.12) имеет как бы «арифметический» характер, является результатом численного совпадения. Качественные отличия взаимосвязей событий выявляются при нарушении симметрии монет. Для несимметричных монет (с вероятностью выпадения герба $\neq 1/2$) свойство независимости A и B вида (1.11) сохраняется, а вот равенства

$$P(AC) = P(A)P(C) \quad \text{и} \quad P(BC) = P(B)P(C)$$

нарушаются.

Тем не менее именно «арифметическое» понимание (1.11) определяет независимость в теории вероятности.

В применении к случайному вектору с независимыми компонентами $X = \{X_1, X_2\}$ это дает

$$\mathbf{P}(X_1 < x_1, X_2 < x_2) = \mathbf{P}(X_1 < x_1)\mathbf{P}(X_2 < x_2),$$

что влечет за собой

$$\boxed{F(x_1, x_2) = F_1(x_1)F_2(x_2),}$$

и, как следствие,

$$\boxed{\rho(x_1, x_2) = \rho_1(x_1)\rho_2(x_2).}$$

Функции $F(x_1, x_2)$ и $\rho(x_1, x_2)$ называются *совместными*, соответственно, функциями и плотностями распределения случайных величин X_1 и X_2. Таким образом, *если случайные величины независимы, то их совместная плотность (функция) распределения равна произведению плотностей (функций)*. Это правило действует и в общем случае n случайных величин — и принимается за определение независимости.

Исходное определение n независимых событий имеет вид

$$\mathbf{P}(A_1 \dots A_n) = \mathbf{P}(A_1) \dots \mathbf{P}(A_n).$$

В сценарии парадокса Бернштейна в случае (1.12) $\mathbf{P}(ABC) = 0$ при ненулевых вероятностях событий A, B, C, откуда ясно, что из *попарной* независимости A, B, C — их независимость не следует.

Обратно. *Возможна независимость при отсутствии попарной независимости.* Соответствующий пример дает бросание двух упорядоченных костей (красной и синей) [31]:

$$A = \{(i, j)\colon\ 1, 2 \text{ или } 5\},$$

$$B = \{(i, j)\colon\ 4, 5 \text{ или } 6\},$$

$$C = \{(i, j)\colon\ i + j = 9\}.$$

Примеры подобного сорта свидетельствуют о наличии подводных течений, но на практике независимость обычно хорошо работает, минуя аномалии. Это тем более справедливо в отношении случайных величин. Там накладывается требование независимости не на два-три события, а на любые комбинации неравенств, что исключает неприятности.

1.8. Дисперсия и ковариация

Скаляр

$$\boxed{\mathbf{D}(X) = \mathbf{E}(X - m_x)^2}$$

называется *дисперсией* случайной величины X, а

$$\sigma_x = \sqrt{\mathbf{D}(X)}$$

— *среднеквадратическим отклонением* X от своего среднего значения m_x.

В силу линейности оператора $\mathbf{E}$:

$$\mathbf{E}(X - m_x)^2 = \mathbf{E}(X^2) - 2\mathbf{E}(X)m_x + m_x^2 = \mathbf{E}(X^2) - m_x^2.$$

Поэтому дисперсия $\mathbf{D}(X)$ равна разности $\mathbf{E}(X^2) - m_x^2$, где $\mathbf{E}(X^2)$ так называемый *второй момент*. Вообще, $\mathbf{E}(X^n)$ именуется *моментом n-го порядка случайной величины X*, в соответствии с чем *матожидание — первый момент*.

Случайная величина $X - m_x$ имеет нулевое матожидание, и ее называют *центрированной*, а моменты центрированных величин — *центральными*. По этой терминологии дисперсия — второй центральный момент.

Из существования $\mathbf{E}(X^n)$ вытекает существование $\mathbf{E}(X^k)$ при любом $k \leqslant n$, причем $\mathbf{E}(X^k) \leqslant [\mathbf{E}(X^n)]^{k/n}$. (?)

Для двух случайных величин X, Y рассматривают *смешанные моменты* $\mathbf{E}(X^n Y^m)$. Важную роль во многих ситуациях играет *ковариация*

$$\boxed{\operatorname{cov}(XY) = \mathbf{E}\left[(X - m_x)(Y - m_y)\right]}$$

и *коэффициент корреляции*

$$r_{xy} = \frac{\operatorname{cov}(XY)}{\sigma_x \sigma_y}.$$

Очевидно,

$$\boxed{\operatorname{cov}(XY) = \mathbf{E}(XY) - m_x m_y,}$$

и $\operatorname{cov}(XY) = 0$, если X и Y независимы. Но ковариация может быть нулевой в случае зависимых X, Y.

Решим, например, такую задачу, считая X, Y — центрированными (для простоты). Найдем приближение Y случайной величиной $Z = \alpha X$ по квадратичному критерию:

$$\mathbf{E}(Y - \alpha X)^2 \to \min. \tag{1.13}$$

Приравнивая нулю производную (1.13) по α, получаем

$$2\mathbf{E}\{(Y - \alpha X)X\} = 0,$$

откуда

$$\alpha = \frac{\operatorname{cov}(XY)}{\mathbf{D}(X)},$$

т. е. при ненулевой ковариации (корреляции) между X и Y существует «линейная зависимость» вида $Y = \alpha X + W$ с ненулевым коэффициентом α и случайной величиной W, некоррелированной с X, $\operatorname{cov}(XW) = 0$.

Практическое вычисление корреляций часто приводило к обнаружению «неожиданных» связей мистического толка. При этом упускалось из вида, что причинная связь и функциональная — совсем разные вещи. Например, процессы, подверженные влиянию солнечной активности, в результате могут коррелировать друг с другом, а их функциональная связь может быть использована для прогноза, но не для объяснения.

Пример. Случайные величины X и $Y=X^2$, при равномерном распределении X в промежутке $[-1, 1]$, — связаны жесткой функциональной зависимостью, но их ковариация равна нулю,

$$\operatorname{cov}(XY) = \int_{-1}^{1} \frac{x(x^2 - m_y)}{2}\, dx = 0,$$

поскольку линейная составляющая взаимосвязи отсутствует.

Некоторые учебники от термина «ковариация» вообще отказываются, заменяя его ***корреляцией*** *или* ***корреляционным моментом***

$$R_{xy} = \operatorname{cov}(XY)$$

и называя коэффициентом корреляции ту же «нормированную» величину

$$r_{xy} = \frac{R_{xy}}{(\sigma_x \sigma_y)}.$$

Это имеет свои минусы, но разгружает терминологию, и выглядит приемлемо.

Упражнения

- Если случайная величина X принимает значения только из интервала $(0, 1)$, то $\sigma_x < m_x$. (?)
- $(\sigma_x - \sigma_y)^2 \leqslant \sigma_{x+y} \leqslant (\sigma_x + \sigma_y)^2$. (?)
- $\sigma_x \cdot \sigma_y \leqslant \sigma_{x \cdot y}$. (?)
- $\sigma_x^2 = \min\limits_{\alpha} \mathbf{E}\{[X - \alpha]^2\}$. (?)

1.9. Неравенства

Неравенство Коши—Буняковского [17]**:**

$$\boxed{\mathbf{E}(|XY|) \leqslant \sqrt{\mathbf{E}(X^2)\mathbf{E}(Y^2)}.} \qquad (1.14)$$

[17] Альтернативные названия: *неравенство Шварца* либо *Коши—Шварца*.

◄ Из $\mathbf{E}\{(\lambda|X| - |Y|)^2\} \geqslant 0$ следует

$$\lambda^2\mathbf{E}(X^2) - 2\lambda\mathbf{E}(|XY|) + \mathbf{E}(Y^2) \geqslant 0,$$

а положительность квадратного многочлена (от λ) влечет за собой отрицательность дискриминанта, что представляет собой доказываемое неравенство. ►

Из (1.14) сразу вытекает, что *коэффициент корреляции всегда по модулю меньше или равен единице.*

Если $\varphi(x) \geqslant 0$ — неубывающая при $x \geqslant a$ функция, то

$$\int_{-\infty}^{\infty} \varphi\, dF(x) \geqslant \int_{a}^{\infty} \varphi\, dF(x) \geqslant \varphi(a) \int_{a}^{\infty} dF(x) = \varphi(a)\mathbf{P}(X \geqslant a),$$

откуда

$$\mathbf{P}(X \geqslant a) \leqslant \frac{\mathbf{E}(\varphi(X))}{\varphi(a)} \quad \text{при условии} \quad \varphi(a) \neq 0. \tag{1.15}$$

Выбор $\varphi(x) = x^2$ и $|X - m_x|$ в качестве случайной величины дает **неравенство Чебышева**:

$$\boxed{\mathbf{P}(|X - m_x| \geqslant a) \leqslant \frac{\mathbf{D}(X)}{a^2}.} \tag{1.16}$$

Из (1.16) следует, что оценка сверху среднеквадратического отклонения влечет за собой оценку сверху вероятности отклонения. Это позволяет переводить разговор из одной плоскости в другую — от моментов к вероятностям.

Так сложилось, что (1.16) затмило другие возможности. Имеет смысл держать в памяти общее неравенство (1.15), из которого можно извлекать более подходящие следствия для конкретных задач. Например, *неравенство Маркова*

$$\mathbf{P}(X > a) \leqslant \frac{\mathbf{E}(X)}{a} \quad \text{при условии} \quad X \geqslant 0$$

или

$$\mathbf{P}(|X| \geqslant \varepsilon) \leqslant \frac{\mathbf{E}(|X|^k)}{\varepsilon^k}, \qquad \mathbf{P}(|X| \geqslant \varepsilon) \leqslant e^{-\alpha\varepsilon}\mathbf{E}(e^{\alpha|X|}) \qquad (k, \varepsilon, \alpha > 0).$$

Неравенство Колмогорова. *Пусть последовательность независимых случайных величин* X_j *имеет нулевые матожидания* $\mathbf{E}(X_j) = 0$ *и* $\mathbf{D}(X_j) < \infty$. *Тогда*

$$\boxed{\mathbf{P}\left\{\max_{k\leqslant n}|X_1+\ldots+X_k|\geqslant\varepsilon\right\}\leqslant\frac{1}{\varepsilon^2}\sum_{j=1}^{n}\mathbf{D}(X_j).} \tag{1.17}$$

◀ Пусть S_k обозначает сумму $X_1 + \ldots + X_k$; A_j — событие, состоящее в том, что

$$|S_j| \geqslant \varepsilon, \quad \text{но} \quad |S_i| < \varepsilon \quad \text{при всех} \quad i < j.$$

Объединение непересекающихся событий A_j есть событие A, означающее $\max\{|S_i| \geqslant \varepsilon,\ i \leqslant n\}$.

Условная дисперсия

$$\mathbf{E}(S_n^2|A_j) = \mathbf{E}\left[(S_n - S_j + S_j)^2|A_j\right] =$$
$$= \mathbf{E}\left[(S_n - S_j)^2|A_j\right] + 2\mathbf{E}\left[(S_n - S_j)S_j|A_j\right] + \mathbf{E}(S_j^2|A_j) \geqslant \varepsilon^2,$$

поскольку $\mathbf{E}\left[(S_n - S_j)^2|A_j\right] \geqslant 0$, второе слагаемое $\mathbf{E}\left[(S_n - S_j)S_j|A_j\right] = 0$, так как $S_n - S_j$ и S_j независимы, потому что состоят из разных независимых слагаемых, а $\mathbf{E}(S_j^2|A_j) \geqslant \varepsilon^2$ — по определению A_j.

Поэтому

$$\mathbf{E}(S_n^2) = \sum_{j=1}^{n}\mathbf{E}(S_n^2|A_j)\mathbf{P}(A_j) + \mathbf{E}(S_n^2|\bar{A})\mathbf{P}(\bar{A}) \geqslant \varepsilon^2\sum_{j=1}^{n}\mathbf{P}(A_j),$$

что и есть (1.17). ▶

Если бы максимум в (1.17) достигался при $k = n$, неравенство сводилось бы к неравенству Чебышева. Из (1.17), разумеется, следует

$$\mathbf{P}\{|S_k| \leqslant \varepsilon;\ k = 1, \ldots, n\} \geqslant 1 - \frac{1}{\varepsilon^2}\sum_{j=1}^{n}\mathbf{D}(X_j).$$

(!) Требование независимости с. в. X_j в (1.17) можно ослабить до

$$\mathbf{E}(X_j|X_1, \ldots, X_{j-1}) = 0 \tag{1.18}$$

при любом j, т. е. заменив независимость предположением о равенстве условных матожиданий безусловным. Обоснование несложно. Независимость X_j при доказательстве неравенства (1.17) использовалась в двух пунктах. При обосновании

$$\mathbf{E}\left[(S_n - S_j)S_j|A_j\right] = 0 \quad \text{и} \quad \mathbf{E}(S_n^2) = \sum_{j=1}^{n}\mathbf{D}(X_j).$$

То и другое остается справедливым без предположения независимости X_j, но при условии (1.18).

Неравенство Иенсена. *Пусть* $\varphi(x)$ — *вогнутая функция* (*выпуклая вверх*) *и матожидание* $\mathbf{E}(X)$ *существует. Тогда*

$$\boxed{\mathbf{E}\,\varphi(X) \leqslant \varphi(\mathbf{E}\,X).} \tag{1.19}$$

◀ Для выпуклой вверх функции $\varphi(x)$ всегда найдется функция[18] $\psi(x)$ такая, что

$$\varphi(x) \leqslant \varphi(y) + \psi(y)(x - y).$$

Матожидание этого неравенства при $x = X$, $y = \mathbf{E}\,X$ дает (1.19). ▶

1.10. Случайные векторы

Если компоненты случайного вектора $X = \{X_1, \ldots, X_n\}$ независимы, то X — просто набор несвязанных друг с другом величин[19].

«Линейная часть взаимосвязей» улавливается ковариациями

$$k_{ij} = \mathbf{E}\left\{(X_i - m_i)(X_j - m_j)\right\} \qquad (m_i = \mathbf{E}(X_i)),$$

которые объединяются в *ковариационную матрицу* $K = [k_{ij}]$.

Корреляционная матрица R из K получается переходом к элементам

$$r_{ij} = \frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}}, \qquad k_{ii} = \mathbf{D}(X_i),$$

т. е. к коэффициентам корреляции.

Обе матрицы K *и* R *неотрицательно определены*, поскольку

$$\sum_{ij} k_{ij}\xi_i\xi_j = \mathbf{E}\left\{\sum_{ij}(X_i - m_i)\xi_i\right\}^2 \geqslant 0.$$

Аналогично для R.

Метод наименьших квадратов. Допустим, на вход объекта (рис. 1.3) действует случайный вектор $X = \{X_1, \ldots, X_n\}$. Скалярный выход Y не определяется входом X, поскольку еще действует ненаблюдаемое возмущение ζ.

---

[18] В гладком случае $\psi(x) = \varphi'(x)$.

[19] По крайней мере, в вероятностном смысле.

Задача состоит в построении линейной модели [20)]

$$Z = \sum_i c_i X_i$$

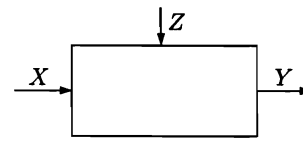


Рис. 1.3

по критерию минимума среднеквадратической ошибки:

$$\mathbf{E}\left(Y - \sum_i c_i X_i\right)^2 \to \min_c .$$

◄ Минимум определяет равенство нулю производных по c_j:

$$\frac{\partial}{\partial c_j}\mathbf{E}\left(Y - \sum_i c_i X_i\right)^2 = 2\mathbf{E}\left\{\left(Y - \sum_i c_i X_i\right) X_j\right\} = 0.$$

Оптимальный вектор c, таким образом, является решением системы

$$K_x c = K_{xy},$$

где K_x ковариационная матрица X, а вектор ковариации K_{xy} имеет координаты $\mathbf{E}\{X_j Y\}$. ►

Обратим внимание, что задача по форме полностью совпадает с поиском ближайшей к вектору y точки $\sum_i c_i x_i$, лежащей в плоскости, натянутой на векторы $\{x_1, \ldots, x_n\}$. Решение последней — как известно — дает ситуация, в которой вектор $y - \sum_i c_i x_i$ ортогонален $\{x_1, \ldots, x_n\}$, т. е. все скалярные произведения $\left(y - \sum_i c_i x_i\right) x_j$ равны нулю.

Это дает естественные основания считать случайные величины с нулевой ковариацией — *ортогональными*. Никакой особой глубины за этим нет, кроме возможности мыслить о с. в. в терминах евклидовых пространств, что иногда раздвигает горизонты.

Модель $Z = \sum_i c_i X_i$ обычно служит для прогноза Y, что, в свою очередь, является основой для принятия экономических или

[20)] В предположении центрированности всех случайных величин.

технологических решений [21]. Проблему поиска оптимальных моделей в рассмотренной постановке называют *задачей идентификации.*

Если линейное преобразование $U = AX$ преобразует случайный вектор X, то, как легко проверить, ковариационные матрицы K_u и K_x связаны соотношением [22]

$$K_u = AK_xA^T, \tag{1.20}$$

где A^T — транспонированная матрица.

Равенство (1.20) есть правило преобразования квадратичной формы при переходе к другому базису [23]. Поскольку квадратичная форма всегда приводится ортогональным преобразованием A к диагональной форме, то от исходного вектора X всегда можно перейти к случайному вектору $U = AX$ с некоррелированными компонентами.

1.11. Вероятностные алгоритмы

Добавление вероятностного фактора в детерминированные задачи иногда преображает ситуацию и дает существенный выигрыш. Суть дела проще всего пояснить на примере.

Проблема выяснения простоты числа N считается пока переборной задачей. Малая теорема Ферма для любого простого N гарантирует равенство

$$a^{N-1} = 1(\mathrm{mod}\, N), \qquad a < N. \tag{1.21}$$

Нарушение (1.21) означает, что N — составное. Но для некоторых составных чисел [24] условие (1.21) тоже выполняется, и поэтому малая теорема Ферма — не очень хорошая лакмусовая бумажка для различения простых и составных чисел. Однако возможна замена (1.21) неким близким условием [25], нарушение которого хотя бы для

---

[21] Y может быть, например, котировкой акций либо параметром, характеризующим эффективность работы химического реактора, прокатного стана и т. п.

[22] $\mathbf{E}\{U_iU_j\} = \mathbf{E}\left\{\sum_{p,q} a_{ip}a_{jq}x_px_q\right\}$.

[23] См., например, [5, т. 3].

[24] Для так называемых чисел Кармайкла.

[25] Подробности и дополнительные ссылки можно найти в: *Нестеренко Ю. В.* Алгоритмические проблемы теории чисел // Матем. просв. 1998. **3**, вып. 2. С. 87–114.

одного $a < N$ гарантирует, что N — составное. Причем для любого составного N подходящих чисел $a < N$ существует не менее $\frac{3}{4}(N-1)$.

При случайном выборе $a < N$, таким образом, составное N не классифицируется как составное с вероятностью не большей $1/4$, а после k проверок — с вероятностью не большей $1/4^k$. После 100 проверок вероятность ошибочной классификации числа N имеет порядок 10^{-60}.

Коллекция задач подобного сорта на сегодняшний день не так велика. Основные трудности заключаются в поиске удобных признаков типа (1.21) для решения альтернативных вопросов. Известно много необходимых условий различного рода, но они в чистом виде, как правило, не годятся — по той же причине, что и (1.21) в рассмотренной задаче.

В целом включение в поле зрения численных методов — вероятностных алгоритмов — породило довольно интересную область исследований. При этом на описанной выше схеме здесь дело не зацикливается. По поводу идеологического разнообразия можно упомянуть «нашумевшую» PCP-теорему. Вольная трактовка ее примерно такова. Существует способ записи математических доказательств, при котором проверка их правильности сводится к анализу нескольких случайно выбранных мест, число которых не зависит от длины исходного доказательства. Поверить, конечно, трудно.

1.12. Об истоках

В ТВ важную роль играют вопросы обоснования, что нередко уводит изложение из сферы интересов широкой аудитории. Нижеследующий текст призван снять некоторую долю напряжения, возникающего в связи с употреблением терминов типа σ-алгебры.

При обобщении исходной вероятностной модели с конечным множеством Ω на счетные и континуальные варианты Ω возникают проблемы. Если в конечном случае можно без предосторожностей рассматривать множество всех подмножеств Ω, то для бесконечных множеств это уже не так.

Необходимость же договориться о том, какие подмножества Ω попадают в поле зрения, возникает из-за того, что сумма и пересечение событий не должны выводить за рамки дозволенного. Но тогда приходится требовать

$$A, B \subset \Omega \quad \Rightarrow \quad A \cup B \subset \Omega, \quad A \cap B \subset \Omega, \tag{1.22}$$

и рассматривать совокупность $\mathcal{A}$ подмножеств Ω, куда входят: само Ω, любое A принадлежит $\mathcal{A}$ вместе с дополнением, — и выполняется (1.22). Такая совокупность $\mathcal{A}$ множеств называется *алгеброй* подмножеств Ω, и — *σ-алгеброй* в более общем случае, когда в $\mathcal{A}$ входят любые суммы и пересечения *счетных совокупностей* $A_k \subset \mathcal{A}$:

$$\bigcup_k^{\infty} A_k \subset \Omega, \qquad \bigcap_k^{\infty} A_k \subset \Omega.$$

Термин «σ-алгебра» обладает способностью отпугивать. Но здесь, в крайнем случае, можно закрыть глаза. Понятия σ-алгебры и меры Лебега (см. далее) — это внутренняя кухня **ТВ**, *юридическая часть. Как бы лицензия на право выполнения различных манипуляций. Если речь идет об аппаратной стороне дела, то о σ-алгебрах можно забыть. Точно так же никто не помнит о дедекиндовых сечениях, лицензирующих использование вещественных чисел.*

Далее возникает проблема задания вероятностей на Ω. В континуальном варианте приходится задавать не $\mathbf{P}(\omega)$, а вероятности кубиков, например. Затем все происходит по схеме определения интегрирования. Сложные фигуры (события) аппроксимируются совокупностями все более мелких кубиков, и пределы объявляются интегралами:

$$\mathbf{P}(A) = \int_A \mu(\omega)\, d\omega \quad \text{либо} \quad \mathbf{E}(X) = \int_{\Omega} X(\omega)\mu(\omega)\, d\omega.$$

В итоге *вероятностным пространством* называют непустое множество Ω с «узаконенным» семейством $\mathcal{A}$ его подмножеств и неотрицательной функцией (мерой) $\mathbf{P}$, определенной на $\mathcal{A}$ и удовлетворяющей условию $\mathbf{P}(\Omega) = 1$, а также

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

для любой последовательности $A_1, A_2, \ldots \in \mathcal{A}$ взаимно непересекающихся множеств A_i. Другими словами, вероятностное пространство определяет тройка $(\Omega, \mathcal{A}, \mathbf{P})$.

Подмножества из $\mathcal{A}$ называют *событиями*.

В зависимости от используемых схем предельных переходов могут получаться разные интегралы — Римана и Лебега. Интегрирование по Риману плохо тем, что чуть что — перестает работать. Скажем, не интегрируются пределы функций. На финише, правда, такого почти никогда не бывает, но многие доказательства рассыпаются. Присказка «интегрируя по Лебегу» обычно спасает положение, потому что по Лебегу интегрируется почти все. Чтобы подчеркнуть отличия, интеграл по Лебегу записывают несколько иначе. Например, так

$$\mathbf{E}(X) = \int_{\Omega} X(\omega)\mu(d\omega).$$

Интегралы Лебега и Римана совпадают, если оба существуют. Поэтому интегрирование простых функций ничем не отличается от обычного, а при интегрировании сложных — до вычислений дело не доходит. Принципиальную важность имеет сама возможность интегрирования по Лебегу. Это сводит концы с концами. Примерно как иррациональные числа. В приближенных вычислениях они не используются, однако, заделывая бреши, превращают вещественную прямую в нормальное игровое поле. Но если о дедекиндовых сечениях при этом можно даже не упоминать, то в **ТВ** иногда требуется умение произносить фразу «интегрируя по Лебегу», не испытывая особого дискомфорта.

Множество Ω с заданной на нем σ-алгеброй $\mathcal{A}$ называют *измеримым пространством*. В случае, когда Ω представляет собой вещественную прямую, — *борелевская σ-алгебра* $\mathcal{B}$ порождается[26] системой непересекающихся полуинтервалов $(\alpha, \beta]$. Элементы $\mathcal{B}$ называют *борелевскими множествами*.

В случае $\Omega = R^n$ борелевские множества определяются аналогично (как прямые произведения одномерных).

Вещественная функция $f(\omega)$ называется *измеримой* относительно σ-алгебры $\mathcal{A}$, если прообраз любого борелевского множества принадлежит $\mathcal{A}$. Если $\mathcal{A} = \mathcal{B}$, функция $f(\omega)$ называется *борелевской*.

[26] Взятием всевозможных объединений и пересечений.

1.13. Задачи и дополнения

- **Парадокс де Мере** связан с бросанием двух игральных костей. Вероятность выпадения двух троек в k бросаниях равна, очевидно, $p_k = 1 - (35/36)^k$. Во времена Блеза Паскаля (XVII в.) вычисление даже такого простого выражения при больших k было обременительно, и оценки часто делались на основе правдоподобных рассуждений. «Одна тройка в 4 бросаниях выпадает с вероятностью $> 1/2$. Две тройки одновременно выпадают в 6 раз реже, чем одна. Поэтому в $24 = 4 \times 6$ бросаниях естественно ожидать $p_{24} > 1/2$». На самом деле $p_{24} \approx 0{,}49$.
  Логические ошибки подобного рода довольно широко распространены на бытовом уровне. «Если лотерейный билет выигрывает с вероятностью p, то 2 билета — с вероятностью $2p$». При малых p в первом приближении это действительно верно. Однако хотя бы один герб при двух бросаниях выпадает с вероятностью $3/4$, а не $2 \cdot (1/2) = 1$.

- *Двумерное неравенство Чебышева*:

$$\mathbf{P}\Big\{\{|X - m_x| \geqslant \varepsilon\sigma_x\} \cup \{|Y - m_y| \geqslant \varepsilon\sigma_y\}\Big\} \leqslant \frac{1 + \sqrt{1 - r_{xy}^2}}{\varepsilon^2}.$$

- Из обычной колоды вытаскивается карта. Если «пика» — это событие A, если «туз» — B. Легко проверяется, что A и B независимы. Но если в колоде есть джокер, — то это не так. (?)

- По поводу избытка парадоксов, заметим следующее. При изучении случайных величин играют роль два фактора: вероятности и значения X. Сознание же не приспособлено следить за двумя параметрами одновременно. В результате простейшие вопросы ставят в тупик.
  Допустим, $\mathbf{P}\{X \leqslant 0\} = \mathbf{P}\{Y \leqslant 0\} \geqslant 1/2$, *причем* X *и* Y *независимы. Вытекает ли отсюда* $\mathbf{P}\{X + Y \leqslant 0\} \geqslant 1/2$?
  Нет. Если X, Y независимо принимают значения $\{-1, 2\}$ с вероятностями $1/2$, то $\mathbf{P}\{X + Y \leqslant 0\} = 1/4$.

- В **ТВ**, как и вообще в жизни, приходится решать нечетко поставленные задачи.
  По равновероятной выборке k чисел из N первых по счету — найти неизвестное N. Чем-то напоминает «угадать фамилию по возрасту», — но задача, вообще говоря, осмысленная.
  Если с. в. X — равна наибольшему числу в выборке, то

$$\mathbf{P}\{X = x\} = \left(\frac{x}{N}\right)^k - \left(\frac{x-1}{N}\right)^k,$$

  откуда при достаточно больших N

$$\mathbf{E}\{X\} \approx \frac{Nk}{k+1}, \qquad \mathbf{D}\{X\} \approx \frac{N^2 k}{(k+1)^2(k+2)}.$$

  Поэтому N оценивается величиной $\dfrac{k+1}{k}X$. Детали уточняются в рамках идеологии сходимости (глава 4).

- **Задача о выборе невесты** в миниатюре служит образцом задач об оптимальных правилах остановки. Сценарий выглядит так. Потенциальному жениху приводят последовательно n девушек. В любой момент он может остановиться: «вот моя невеста», — но возможности вернуться к какому-либо предыдущему варианту нет.
  Как гантели полезны для упражнений, но не для созерцания, — так и эта задача. Думать можно над эквивалентным вариантом: последовательно просматривая числа
  $$\xi_1, \ldots, \xi_n,$$
  в какой-то момент надо остановиться и выбрать ξ_k (как можно большее).
  Среди стратегий «просматриваются первые m чисел, после чего выбирается первое же, превосходящее все $\xi_1, \ldots, \xi_m$» — максимальную вероятность выбрать наибольшее ξ_k дает m, ближайшее к n/e. (?)

- Простота базовой модели теории вероятностей (пространство элементарных событий Ω с заданными на нем вероятностями) нелегко далась исторически, и она нелегко достигается по сей день, ибо многие задачи к каноническому виду сводятся с большим трудом. Это, конечно, не удивительно. Очень простая схема, но в нее укладывается все разнообразие вероятностных задач. Только «укладывание» требует иногда большой изобретательности. Поэтому для освоения **ТВ** необходимо развитие навыков решения задач. В хаосе разнообразных идей и технических приемов здесь есть наезженные пути и характерные модели. Определенный интерес в этом отношении представляет **метод фиктивного погружения.**

  Рассмо́трим *парадокс раздела ставки*[27].
  Матч до 6 *побед прекращен досрочно при счете* 5:3. *В какой пропорции разделить приз?*
  Конечно, это не парадокс, а проблема. Проблема, а не задача, потому что вопрос надо еще правильно поставить. Наиболее логичен был Ферма.

  ◀ Его идея — в гипотетическом *продолжении* игры тремя фиктивными партиями (даже если некоторые из них окажутся лишними). При равновероятности всех 8 исходов второй игрок выигрывает матч лишь в одном случае, — если побеждает во всех трех партиях, — поэтому справедливая пропорция 7 : 1. ▶

- Погружение задачи в более широкий круг фиктивных ситуаций во многих случаях дает выход из положения либо обеспечивает дополнительные удобства. Рассмотрим, для примера, *задачу Банаха.*
  В двух коробках имеется по n спичек. На каждом шаге наугад выбирается коробка, и из нее удаляется одна спичка. Найти вероятность p_k того, что в момент окончания процесса, т. е. опустошения одной из коробок, в другой — остается k спичек.

[27] Об исторических подробностях см. [22].

◀ Если одна коробка пуста, а в другой — k спичек, это означает, что спички брались $2n-k$ раз, причем n раз из (теперь уже) пустой коробки. Поэтому $p_k = C_{2n-k}^{n}/2^{2n-k}$. ▶

При необходимости изучать задачу в целом (распределение p_k при разных k) возникает определенное неудобство, связанное с выбором пространства элементарных событий Ω. Вариант опустошения одной из коробок в момент $n+j$ происходит на фоне других вариантов, которые *из-за переменной длины имеют разные вероятности*. В итоге получается порочный круг. Для решения задачи надо построить Ω, а для построения Ω требуется указать вероятности, которые ищутся. Узел развязывает добавление к настоящим — фиктивных спичек. Тогда в качестве Ω можно рассматривать 2^{2n+1} *равновероятных* вариантов длины $2n+1$. Такой длины всегда хватает для опустошения одной из коробок.

Глава 2

Функции распределения

2.1. Основные стандарты

Равномерное распределение в промежутке $[a, b]$ имеет плотность

$$\rho(x) = \frac{1}{b-a},$$

которой соответствует функция распределения

$$F(x) = \int_{-\infty}^{x} \rho(u)\,du = \frac{1}{b-a}\int_{a}^{x} du = \frac{x-a}{b-a}$$

при $x \in [a, b]$. Разумеется, $F(x) = 0$ при $x \leqslant a$ и $F(x) = 1$ при $x \geqslant b$.

Биномиальное распределение. Среди эталонных вероятностных моделей особое место занимает схема бросания монеты, порождающая цепочки «герб–решетка»: ГРГГР... Если при выпадении герба писать единицу, решетки — нуль, модель будет генерировать случайные «01»-последовательности:

$$1\ 0\ 1\ 1\ 0\ \ldots$$

При этом можно говорить о генерации двоичных чисел вида $0{,}10110\ldots$.

В общем случае в результате испытания (бросания, эксперимента) единица появляется с вероятностью $p \in (0, 1)$, нуль — с вероятностью $q = 1 - p$. Появление единицы часто именуют *успехом*. Проведение соответствующих независимых испытаний называют *схемой*, или *последовательностью испытаний Бернулли*.

В силу независимости испытаний вероятности появления 1 или 0 перемножаются. Поэтому вероятность в n испытаниях получить k единиц в каком-либо определенном порядке (и, соответственно, $n - k$ нулей) — равна $p^k q^{n-k}$. А поскольку k единиц расположить

в n разрядах можно числом способов C_n^k, то вероятность получить k единиц независимо от порядка их следования — равна

$$\boxed{p_k = C_n^k p^k q^{n-k}.}$$

Набор таких вероятностей $\{p_0, \ldots, p_n\}$ называют *биномиальным распределением* (в серии испытаний длины n). Можно сказать, что биномиальное распределение имеет сумма

$$S_n = X_1 + \ldots + X_n,$$

где все с. в. X_k независимы и принимают два возможных значения 1 или 0 с вероятностями p и $q = 1 - p$.

Легко проверить:

$$\mathbf{E}\{S_n\} = np, \qquad \mathbf{D}\{S_n\} = np(1-p), \qquad \mathbf{E}\{[S_n - np]^3\} = np(1-p)(1-2p).$$

На базе бросания монеты часто говорят об *игре в «орлянку»*: герб — выиграл, решетка — проиграл. При этом удобно считать, что X_k принимают значения не 1 и 0, а 1 и -1. За этой схемой, в свою очередь, подразумевают иногда *случайное блуждание* частицы (или выигрыша).

Геометрическое распределение. В схеме Бернулли вероятность появления k нулей перед первым появлением единицы, очевидно, равна $\boxed{p_k = pq^k}$. Совокупность этих вероятностей (при $k = 0, 1, 2, \ldots$) называют *геометрическим распределением*[1]. Вероятность первого успеха, соответственно, равна

$$\mathbf{P}\{X = x\} = pq^{x-1}.$$

Несложный подсчет показывает:

$$\mathbf{E}\{X\} = \frac{1}{p}, \qquad \mathbf{D}\{X\} = \frac{q}{p^2}.$$

[1] Геометрическое распределение имеет случайная величина, равная числу испытаний до первого успеха — число промахов до первого попадания, т. е. «число лягушек, которых приходится перецеловать, пока не найдешь своего принца».

В качестве механизма организации последовательных испытаний Бернулли могут использоваться *урновые модели*. В урне находится k белых шаров и m черных. Вероятность вытащить белый шар равна $p = \dfrac{k}{k+m}$, черный — $q = \dfrac{m}{k+m}$. При последовательном извлечении шаров возможны два варианта: шар, вытащенный на предыдущем шаге, возвращается в урну или не возвращается.

Встречаются постановки задачи с бóльшим количеством цветов. По существу, урновой является карточная модель с популярными задачами типа: «из колоды вытаскивается n карт — какова вероятность, что k из них одной масти?».

В основе урновых моделей лежит равновероятный выбор любого из шаров. Симметрию нарушает раскраска. «Сложные» в данном случае события выбора белого или черного шаров можно взять в качестве элементарных — для схемы Бернулли. Это дает готовый механизм обеспечения вероятности $p = \dfrac{k}{k+m}$.

Иногда говорят, что погоду в теории вероятностей определяют три закона распределения: *биномиальный*, *нормальный* и *пуассоновский*. Из дальнейшего будет видно, что из этой тройки два последних можно в некотором роде исключить. Нормальное распределение и пуассоновское являются асимптотическими вариантами — биномиального.

Распределение Пуассона, как и биномиальное, является дискретным, и характеризуется вероятностями

$$\boxed{\mathrm{P}(X = k) = \frac{a^k}{k!}e^{-a}} \qquad (k = 0, 1, \ldots).$$

Легко убедиться, что

$$a = \sum_{k=0}^{\infty} k\mathrm{P}(X = k),$$

т. е. параметр a есть матожидание с. в. X, распределенной по закону Пуассона. Дисперсия X тоже равна a.

Закон $p_k = a^k e^{-a}/k!$ получается из биномиального, если $n \to \infty$ и при этом вероятность p меняется так, что $pn \to a$.

Действительно, $C_n^k p^k (1-p)^{n-k}$ при условии $p = a/n$ можно записать в виде

$$\frac{(np)^k}{k!}(1-p)^n \frac{\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)}{(1-p)^k} = \frac{a^k}{k!}\left[\left(1-\frac{a}{n}\right)^{n/a}\right]^a \frac{\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)}{\left(1-\frac{a}{n}\right)^k}.$$

Закон Пуассона получается с учетом

$$\left(1-\frac{a}{n}\right)^{n/a} \to e^{-1}, \quad \frac{\left(1-\frac{1}{n}\right)\dots\left(1-\frac{k-1}{n}\right)}{\left(1-\frac{a}{n}\right)^k} \to 1 \quad \text{при} \quad n \to \infty.$$

Но генеалогическое древо пуассоновского распределения имеет более важные ответвления (см. раздел 2.8).

Нормальный закон распределения. Случайные величины, с которыми приходится иметь дело на практике, чаще всего подчинены *нормальному закону распределения*[2], имеющему плотности вида

$$\rho(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}. \tag{2.1}$$

Различия определяются матожиданием m_x и дисперсией σ_x^2. Примеры графиков плотностей (2.1) изображены на рис. 2.1.

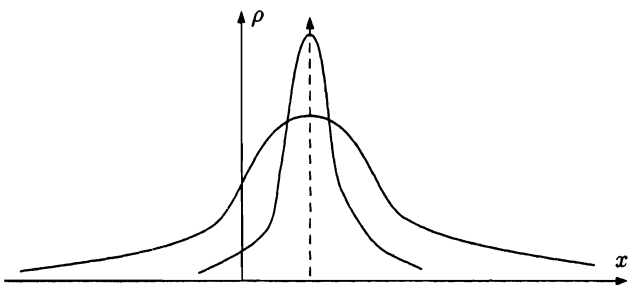


Рис. 2.1

Причины, по которым нормальный закон широко распространен в природе, анализируются в разделе 2.7. Для краткости речи иногда используют обозначение $\mathcal{N}(m_x, \sigma_x^2)$. Например, $\mathcal{N}(0, 1)$ обозначает нормальное распределение с нулевым матожиданием и единичной дисперсией. Функция распределения $\mathcal{N}(0, 1)$ имеет вид

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2}\, ds. \tag{2.2}$$

[2] Нормальное распределение называют также *гауссовским*.

Интеграл (2.2) не выражается через элементарные функции. Вместо «стандарта» $\Phi(x)$ используется также интеграл

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-s^2/2}\, ds,$$

связанный с (2.2) очевидным соотношением $\Psi(x) + 1/2 = \Phi(x)$.

Упражнения

- Если X и Y распределены геометрически, то и $Z = \min\{X, Y\}$ имеет геометрическое распределение. (?)
- Если X и Y распределены нормально либо по Пуассону, то $Z = X + Y$ имеет, соответственно, такое же распределение. (?)
- Если X и Y имеют функции распределения $F_x(x)$ и $F_y(y)$, то с. в. $Z = \max\{X, Y\}$ имеет функцию распределения $F_z(z) = F_x(z)F_y(z)$. (?)

2.2. Дельта-функция

Использование дельта-функций для записи плотностей распределения сводит воедино непрерывные и дискретные задачи и позволяет рассматривать смешанные задачи с плотностями

$$\rho(x) = p_c\rho_c(x) + p_1\delta(x - x_1) + \ldots + p_k\delta(x - x_k).$$

Плотность распределения пуассоновского закона, в частности, имеет вид

$$\rho(x) = \sum_{k=0}^{\infty} \frac{a^k}{k!} e^{-a}\delta(x - k).$$

Такого сорта выгоды иногда трактуются как чисто технические нюансы, способствующие обозримости и единообразию результатов. Здесь можно добавить, что удобства — это, как правило, вопрос жизни и смерти математической дисциплины.

Дельта-функция $\delta(x)$ изначально определялась как предел единичных импульсов[3] $\delta_\varepsilon(x)$, прямоугольной формы либо колоколообразной (рис. 2.2), — при стремлении к нулю ширины импульса, $\varepsilon \to 0$.

[3] Единичной площади, $\int \delta_\varepsilon(x)\, dx = 1$.

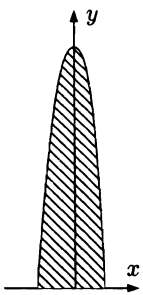


Рис. 2.2

При $\varepsilon \to 0$ никакого разумного предела в обычном смысле нет, но ситуации, в которых возникает такая потребность, обычно сводятся к сходимости интеграла

$$\int_{-\infty}^{\infty} \delta_\varepsilon(x)\varphi(x)\,dx \to \varphi(0) \quad \text{при} \quad \varepsilon \to 0,$$

что и закладывается в понимание предела

$$\delta_\varepsilon(x) \to \delta(x).$$

Такое понимание предельных переходов работает во многих других ситуациях, составляющих базу теории обобщенных функций, где сначала вводится понятие пространства $\mathcal{D}$ *основных функций* — бесконечно дифференцируемых *финитных функций*. Финитных в том смысле, что $\varphi(x) \equiv 0$ вне ограниченной области (не общей для всех, а своей для каждой $\varphi \in \mathcal{D}$).

Обобщенные функции затем определяются как *линейные функционалы* f над $\mathcal{D}$, ставящие любой функции $\varphi \in \mathcal{D}$ в соответствие «скалярное произведение» (f, φ). Простейший пример линейного функционала дает интегральное представление

$$(f, \varphi) = \int_{-\infty}^{\infty} f(x)\varphi(x)\,dx.$$

В соответствии с этой идеологией обобщенная функция $\delta(x)$ — это «нечто», действующее на функции $\varphi \in \mathcal{D}$ по правилу

$$\boxed{(\delta, \varphi) = \int_{-\infty}^{\infty} \delta(x)\varphi(x)\,dx = \varphi(0).}$$

Что касается обычных функций $f(x)$, то они одновременно — и обобщенные, действующие на $\varphi \in \mathcal{D}$ в рамках определения скалярного произведения

$$(f, \varphi) = \int f(x)\varphi(x)\,dx.$$

Производные обобщенных функций определяются равенством

$$\boxed{\int_{-\infty}^{\infty} f'(x)\varphi(x)\,dx = -\int_{-\infty}^{\infty} f(x)\varphi'(x)\,dx,} \tag{2.3}$$

что можно воспринимать как результат интегрирования по частям левого интеграла. Обращение в нуль слагаемого[4] $f(x)\varphi(x)\big|_{-\infty}^{\infty}$ происходит из-за финитности $\varphi(x)$.

Для производной $\delta'(x)$ равенство (2.3) приводит к

$$\int_{-\infty}^{\infty} \delta'(x)\varphi(x)\,dx = \varphi'(0).$$

Следствием (2.3) является также важное соотношение

$$\boxed{\theta'(x) = \delta(x),}$$

где $\theta(x)$ — *функция Хэвисайда*, единичная ступенька:

$$\theta(x) = \begin{cases} 1, & x > 0; \\ 0, & x < 0. \end{cases}$$

◀ Действительно, для любой функции $\varphi \in \mathcal{D}$

$$\int_{-\infty}^{\infty} \theta'(x)\varphi(x)\,dx = -\int_{-\infty}^{\infty} \theta(x)\varphi'(x)\,dx = -\int_{0}^{\infty} \varphi'(x)\,dx = \varphi(0),$$

т. е. производная $\theta'(x)$ действует на φ так же, как $\delta(x)$. ▶

Замена переменных при интегрировании приводит к формулам:

$$\int_{-\infty}^{\infty} \delta(x-a)\varphi(x)\,dx = \varphi(a), \qquad \int_{-\infty}^{\infty} \delta(ax)\varphi(x)\,dx = \frac{1}{a}\varphi(0).$$

Отметим, наконец, соотношение

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\lambda x}\,d\lambda,$$

где несобственный интеграл понимается как его главное значение.

2.3. Функции случайных величин

Если $Y = f(X)$, где f — обычная детерминированная функция, а X — случайная величина с плотностью $\rho(x)$, то среднее значение

[4] Возникающего при взятии интеграла по частям.

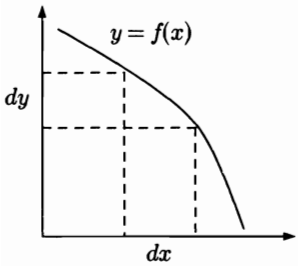


Рис. 2.3

$Y = f(X)$, очевидно, равно[5]

$$m_y = \mathbf{E}(Y) = \int f(x)\rho(x)\,dx.$$

Аналогично,

$$\sigma_y^2 = \mathbf{D}(Y) = \int [f(x) - m_y]^2 \rho(x)\,dx.$$

Если $Y = f(X)$ вектор, подобным образом определяется и ковариация:

$$\operatorname{cov}(Y_i Y_j) = \int [f_i(x) - m_{y_i}][f_j(x) - m_{y_j}]\rho(x)\,dx.$$

С определением плотности распределения $Y = f(X)$ возни немного больше. Из рис. 2.3 видно, что[6]

$$\mathbf{P}\{y < f(X) < y + dy\} = \rho_x\left(f^{-1}(y)\right)\left|\left[f^{-1}(y)\right]'\right||dy|,$$

откуда функция распределения

$$F(y) = \mathbf{P}\{Y < y\} = \int_{-\infty}^{y} \rho_x\left(f^{-1}(y)\right)\left|\left[f^{-1}(y)\right]'\right| dy,$$

а плотность[7] —

$$\rho_y(y) = \rho_x\left(f^{-1}(y)\right)\left|\left[f^{-1}(y)\right]'\right|, \tag{2.4}$$

где индексы x, y показывают, какие плотности подразумеваются.

Если X и Y — векторы, имеющие одинаковую размерность, и

$$X = f^{-1}(Y) = h(Y),$$

то

$$F(y) = \mathbf{P}\{Y < y\} = \int_{-\infty}^{y_1} \dots \int_{-\infty}^{y_n} \rho_x(h(y)) \det\left[\frac{\partial h_i}{\partial y_j}\right] dy,$$

---

[5] Напоминаем, что бесконечные пределы в $\int$ иногда опускаются.

[6] Необходимые оговорки здесь и далее очевидны и для краткости — опущены.

[7] Подразумевается, что $y = f(x)$ имеет единственное решение при любом y. Общий случай рассматривается в следующем разделе.

где

$$\rho_x(h(y))\det\left[\frac{\partial h_i}{\partial y_j}\right] = \rho_y(y),$$

а индексы x, y показывают, какие плотности имеются в виду.

Упражнения

- Если $F(x)$ непрерывная функция распределения с. в. X, то случайная величина $Y = F(X)$ равномерно распределена на $[0, 1]$. (?)
- Пусть F и G непрерывные функции распределения с. в. X и Y. Тогда произведение $Z = XY$ имеет функцией распределения

$$\int_{-\infty}^{0}\left[1 - G\left(\frac{z}{\eta}\right)\right] dF(\eta) + \int_{0}^{\infty} G\left(\frac{z}{\eta}\right) dF(\eta). \quad (?)$$

2.4. Условные плотности

При известной функции распределения

$$F(u, v) = \mathbf{P}\{U < u, V < v\}$$

случайного вектора $X = \{U, V\}$ имеем

$$F_u(u) = \mathbf{P}\{U < u\} = \mathbf{P}\{U < u, V < \infty\} = F(u, \infty).$$

Аналогично,

$$\boxed{F_v(v) = F(\infty, v).}$$

С другой стороны,

$$F_u(u) = \int_{-\infty}^{u}\left\{\int_{-\infty}^{\infty}\rho(u, v)\, dv\right\} du,$$

откуда

$$\boxed{\rho_u(u) = \int_{-\infty}^{\infty}\rho(u, v)dv, \qquad \rho_v(v) = \int_{-\infty}^{\infty}\rho(u, v)\, du.}$$

Условные плотности. Если события A и B означают, соответственно, выполнение неравенств:

$$x < X < x + \Delta x, \qquad y < Y < y + \Delta y,$$

то при достаточно малых Δx и Δy:

$$\mathbf{P}(AB) \approx \rho(x, y)\Delta x \Delta y, \qquad \mathbf{P}(A) \approx \rho(x)\Delta x, \qquad \mathbf{P}(B|A) \approx \rho(y|x)\Delta y.$$

Подставляя эти равенства в формулу $\mathbf{P}(B|A) = \dfrac{\mathbf{P}(AB)}{\mathbf{P}(A)}$ и переходя к пределу при $\Delta x, \Delta y \to 0$, получаем

$$\boxed{\rho(y|x) = \frac{\rho(x, y)}{\rho(x)},} \tag{2.5}$$

что определяет *условную плотность вероятности* $\rho(y|x)$.

Объяснять такие формулы так же легко и сложно, как объяснять, что такое чувство голода. Для кого-то, возможно, легче заменить вероятностную интерпретацию механической. Суть дела от этого не меняется. Пластинка L единичной массы имеет плотность $\rho(x, y)$. Тогда

$$\rho_x(x) = \int_L \rho(x, y)\, dy$$

это плотность распределения массы по x, а $\rho(y|x_0)$ — относительная плотность распределения массы в сечении $x = x_0$. Точнее говоря, плотность распределения в полосе

$$x_0 < x < x_0 + \Delta x$$

при нормировании массы полосы на единицу и $\Delta x \to 0$.
В любом случае с этим стоит немного повозиться, чтобы исходные понятия и тривиальные по сути соотношения не отвлекали при рассмотрении более сложных ситуаций.

Из (2.5) вытекает часто используемая формула

$$\boxed{\rho(x, y) = \rho(y|x)\rho(x).} \tag{2.6}$$

Понятно, что в (2.6) x и y можно поменять местами.

Условные матожидания. Через условную плотность определяются любые условные моменты, в том числе *условное матожидание*:

$$\mathbf{E}(Y|x) = \int y\rho(y|x)\,dy.$$

Условное матожидание представляет собой решение оптимизационной задачи

$$\mathbf{E}[Y - \varphi(X)]^2 \to \min_{\varphi}, \tag{2.7}$$

где минимум ищется по функции φ. Решением оказывается

$$\varphi(x) = \mathbf{E}(Y|x),$$

т. е. $\varphi(X) = \mathbf{E}(Y|X)$ представляет собой наилучшее среднеквадратическое приближение зависимости Y от X, которое называют *регрессией*.

◀ Не углубляясь в детали, поясним сказанное. Приравнивая нулю вариацию

$$\mathbf{E}[Y - \varphi(X)]^2 = \iint [y - \varphi(x)]^2 \rho(x, y)\,dx\,dy,$$

получаем

$$\iint [y - \varphi(x)]\Delta\varphi(x)\rho(y|x)\rho(x)\,dx\,dy = 0,$$

откуда, в силу произвольности вариации $\Delta\varphi(x)$,

$$\varphi(x) = \int y\rho(y|x)\,dy = \mathbf{E}(Y|x). \quad ▶$$

Использование дельта-функций. В случае жесткой функциональной связи $Y = f(X)$ величина Y принимает единственно возможное значение $y = f(x)$, если $X = x$. Поэтому

$$\boxed{\rho(y|x) = \delta[y - f(x)],}$$

что влечет за собой

$$\boxed{\rho(x, y) = \rho_x(x)\delta[y - f(x)]}$$

и, в силу

$$\rho_y(y) = \int \rho(y, x)\,dx,$$

приводит к формуле

$$\boxed{\rho_y(y) = \int \rho_x(x)\delta[y - f(x)]\,dx.} \tag{2.8}$$

Интегрирование (2.8) в точках y, которым соответствуют простые изолированные корни $x_j(y)$ уравнения $y - f(x) = 0$, дает[8)]

$$\rho_y(y) = \sum_j \frac{\rho_x(x_j)}{|f'(x_j)|},$$

что совпадает с (2.4) в случае одного корня $x_j(y)$.

2.5. Характеристические функции

Соотношение (2.8) работает и в ситуации случайных векторов. Для суммы случайных величин $Z = X + Y$, в частности,

$$\rho_z(z) = \iint \rho(x, y)\delta(z - x - y)\,dx\,dy. \tag{2.9}$$

Интегрирование (2.9) по y — дает

$$\rho_z(z) = \int_{-\infty}^{\infty} \rho(x, z - x)\,dx. \tag{2.10}$$

При независимости X и Y (2.10) переходит в

$$\rho_z(z) = \int_{-\infty}^{\infty} \rho_x(x)\rho_y(z - x)\,dx. \tag{2.11}$$

Формула (2.11) представляет собой *свертку плотностей*, в связи с чем в **ТВ** оказываются эффективны[9)] *характеристические функции* (х. ф.) $\boxed{\varphi(\lambda) = \mathbf{E}\,(e^{i\lambda X})}$, т. е.

$$\varphi(\lambda) = \int e^{i\lambda x}\,dF(x),$$

[8)] Независимо от того, конечно или бесконечно число корней. Если мера множества точек, где производная $f'(x)$ вырождена, — равна нулю, то это множество можно просто игнорировать без ущерба для решения задачи.

[9)] О причинах см. далее.

либо

$$\varphi(\lambda) = \int_{-\infty}^{\infty} \rho(x) e^{i\lambda x}\, dx, \qquad i^2 = -1, \qquad -\infty < \lambda < \infty,$$

что в несущественных деталях отличается от стандартного *преобразования Фурье* плотности $\rho(x)$. При условии абсолютной интегрируемости $\int |\varphi(\lambda)|\, d\lambda < \infty$, т. е. $\varphi(\lambda) \in L_1$, соответствующая плотность однозначно восстанавливается «обратным преобразованием Фурье»

$$\rho(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(\lambda) e^{-i\lambda x} d\lambda.$$

Этот факт «территориально» принадлежит другим дисциплинам, но его обоснование, в том числе, можно найти во многих стандартных курсах теории вероятностей.

Если с. в. $X_1, \ldots, X_n$ независимы, то

$$\mathbf{E}\,(e^{i\lambda(X_1+\ldots+X_n)}) = \mathbf{E}\,(e^{i\lambda X_1}) \ldots \mathbf{E}\,(e^{i\lambda X_n}), \tag{2.12}$$

то есть х. ф. $\varphi(\lambda)$ суммы $X_1+\ldots+X_n$ равна произведению х. ф. слагаемых: $\varphi(\lambda) = \prod_k \varphi_k(\lambda)$. Это обстоятельство и определяет заметную роль характеристических функций в теории вероятностей.

Вот характеристические функции стандартных распределений:

распределение	*плотность*	*х. ф.* $\varphi(\lambda)$
нормальное	$\rho(x) = \dfrac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}$	$e^{im_x\lambda - \frac{1}{2}\sigma_x^2\lambda^2}$
равномерное	$\rho(x) = \dfrac{1}{b-a}$ на $[a, b]$	$\dfrac{e^{i\lambda b} - e^{i\lambda a}}{i\lambda(b-a)}$
Коши	$\rho(x) = \dfrac{a}{\pi(x^2+a^2)}$	$e^{-a\lvert\lambda\rvert}$
показательное	$\rho(x) = ae^{-ax}, \quad x \geqslant 0$	$\dfrac{a}{a-i\lambda}$
показательное-2	$\rho(x) = \dfrac{1}{2}e^{-\lvert x\rvert}$	$\dfrac{1}{1+\lambda^2}$

Отметим простейшие свойства х. ф.

- Из $|\mathbf{E}\,(e^{i\lambda X})| \leqslant \mathbf{E}\,|e^{i\lambda X}|$ следует $|\varphi(\lambda)| \leqslant 1$.
- Если $\varphi(\lambda)$ — х. ф. случайной величины X, то $Y = \alpha X + \beta$ имеет характеристическую функцию $e^{i\lambda\beta}\varphi(\alpha\lambda)$.
- Разложение в ряд экспоненты $\varphi(\lambda) = \mathbf{E}\,(e^{i\lambda X})$ приводит к

$$\varphi(\lambda) = \sum_{k=0}^{\infty} \frac{(i\lambda)^k}{k!}\mathbf{E}\,(X^k), \tag{2.13}$$

что — в связи с $\varphi(\lambda) = \sum\limits_{k=0}^{\infty} \varphi^{(k)}(0)\dfrac{(\lambda)^k}{k!}$ — означает

$$\mathbf{E}\,(X^k) = i^{-k}\varphi^{(k)}(0), \tag{2.14}$$

но для этого, конечно, требуется существование моментов $\mathbf{E}\,(X^k)$. Однако, если моменты $\mathbf{E}\,(X^k)$ существуют для $k \leqslant j$, то можно утверждать, что (2.14) справедливо для тех же $k \leqslant j$. (?)

Таким образом, при известной х. ф. определение моментов сводится к простому вычислению производных $\varphi^{(k)}(0)$.

- Вместо характеристических функций иногда удобнее рассматривать их логарифмы. Соответственно, вместо моментов (2.14) — коэффициенты

$$\nu_k = i^{-k}\left.\frac{d^k \ln \varphi(\lambda)}{d\lambda^k}\right|_{\lambda=0},$$

называемые *семиинвариантами*. Семиинварианты суммы независимых с. в. равны суммам семиинвариантов слагаемых. (?) Очевидно, у нормального закона все семиинварианты выше второго порядка равны нулю.

Пример. Пусть независимые с. в. X и Y распределены равномерно, соответственно, на промежутках $[-a, a]$ и $[-b, b]$ $(a < b)$. Тогда $Z = X + Y$ имеет плотность

$$\rho_z(z) = \iint \rho(x, y)\delta(z - x - y)\,dx\,dy = \frac{1}{4ab}\int_{-a}^{a}\int_{-b}^{b} \delta(z - x - y)\,dy\,dx,$$

где $\rho(x, y) = \rho_x(x)\rho_y(y)$.

Интегрирование приводит к функции $\rho_z(z)$, график которой изображен на рис. 2.4. При $a = b$ получается треугольное распределение.

Если $Z = X + Y$ и слагаемые распределены нормально, то перемножение характеристических функций

$$\exp\left\{im_x\lambda - \frac{1}{2}\sigma_x^2\lambda^2\right\}$$

и

$$\exp\left\{im_y\lambda - \frac{1}{2}\sigma_y^2\lambda^2\right\}$$

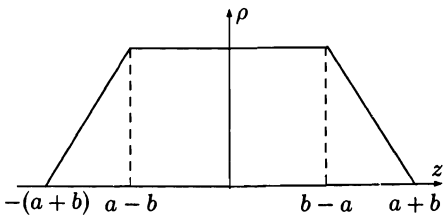


Рис. 2.4

дает характеристическую функцию Z:

$$\exp\left\{i(m_x + m_y)\lambda - \frac{1}{2}(\sigma_x^2 + \sigma_y^2)\lambda^2\right\},$$

откуда видно, что *сумма нормально распределенных с. в. тоже нормально распределена*, причем матожидания и дисперсии просто складываются.

Перемножение характеристических функций сразу дает аналогичный результат, если слагаемые в $Z = X + Y$ имеют распределение Пуассона или Коши. Получение тех же выводов без х. ф. более громоздко.

2.6. Производящие функции

Есть такая задача о взвешивании монет. В одном из 100 *мешков находятся фальшивые монеты. Настоящая монета весит* 7 *грамм, фальшивая —* 6. *Надо с помощью одного взвешивания определить мешок с фальшивыми монетами.*

◀ Мешки нумеруются, после чего из k-го мешка извлекаются k монет, и эти $N = 1 + 2 + \ldots + 100$ монет все вместе взвешиваются. Число недостающих до $7N$ грамм будет номером «фальшивого» мешка. ▶

Вымышленная задача отражает в миниатюре идею, применимую в широком диапазоне различных ситуаций.

Определение. *Производящей функцией числовой последовательности* $a_0, a_1, a_2, \ldots$ *называется ряд*

$$A(z) = \sum_{k=0}^{\infty} a_k z^k.$$

Как совокупность $1 + 2 + \ldots + 100$ монет несла на себе всю информацию о задаче, потому что из каждого мешка было взято разное число монет, — так и производящая функция $A(z)$ несет на себе всю информацию о последовательности $a_0, a_1, a_2, \ldots$, потому что a_k умножаются на z в разных степенях. После такого умножения члены $a_k z^k$ можно безопасно складывать вместе — информация не теряется [10].

Разумеется, эффект от использования производящих функций возникает, если ряд $\sum a_k z^k$ удается свернуть. Широко известна производящая функция $(1 + z)^n = \sum_{k=0}^{n} C_n^k z^k$, порождающая различные связи между биномиальными коэффициентами:

$$z = 1 \quad \Rightarrow \quad \sum_{k=0}^{n} C_n^k = 2^n; \qquad z = -1 \quad \Rightarrow \quad \sum_{k=0}^{n} (-1)^k C_n^k = 0.$$

Более интересные примеры см. в [24].

Если случайная величина X принимает дискретные значения $X = k$ с вероятностями p_k, то

$$\Pi(z) = \sum_{k=0}^{\infty} p_k z^k$$

называют производящей функцией с. в. X. В общем случае целочисленной случайной величины X производящая функция

$$\Pi(z) = \mathrm{E}\{z^X\}.$$

С характеристической функцией $\varphi(\lambda)$ ее связывает соотношение

$$\varphi(\lambda) = \Pi(e^{i\lambda}).$$

При упоминании геометрического распределения иногда имеют в виду число промахов k до первого попадания — и тогда $p_k = pq^k$. А иногда — номер x первого попадания, и тогда $p_x = pq^{x-1}$. В последнем случае

$$\Pi(z) = \mathrm{E}\{z^X\} = \sum_{x=1}^{\infty} pq^{x-1} z^x = \frac{pz}{1 - qz}.$$

[10] Аналогичным образом получаются ряды Фурье.

дискретное распределение	*дискретная плотность*	*х. ф.* $\varphi(\lambda)$	*п. ф.* $\Pi(z)$
биномиальное	$p_k = C_n^k p^k q^{n-k}$	$(p^{i\lambda} + q)^n$	$(pz + q)^n$
геометрическое	$p_k = pq^k$	$\dfrac{p}{1 - qe^{i\lambda}}$	$\dfrac{p}{1 - qz}$
пуассоновское	$p_k = \dfrac{a^k}{k!} e^{-a}$	$e^{a(e^{i\lambda}-1)}$	$e^{-a(1-z)}$

Первые моменты определяются исходя из формул

$$\mathbf{E}\{X\} = \Pi'(1), \qquad \mathbf{E}\{X^2\} - \mathbf{E}\{X\} = \Pi''(1). \tag{2.15}$$

Упражнения

- Для независимых с. в. X и Y

$$\boxed{\Pi_{X+Y}(z) = \Pi_X(z)\Pi_Y(z).} \qquad (?)$$

- Если

$$\Pi(z) = \sum_{k=0}^{\infty} p_k z^k, \qquad \Upsilon(z) = \sum_{k=0}^{\infty} q_k z^k,$$

где $q_k = p_{k+1} + p_{k+2} + \ldots$ — вероятности «хвостов» распределения, то

$$\Upsilon(z) = \frac{1 - \Pi(z)}{1 - z}. \qquad (?)$$

- Если с. в. X имеет распределение $p_1, p_2, \ldots$, то в обозначениях предыдущего пункта:

$$\mathbf{E}\{X\} = \sum_{k=1}^{\infty} k p_k = \sum_{k=1}^{\infty} q_k$$

либо, на языке производящих функций,

$$\mathbf{E}\{X\} = \Pi'(1) = \Upsilon(1).$$

2.7. Нормальный закон распределения

Широкое распространение *нормального закона* [11]

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)}, \tag{2.16}$$

[11] Для простоты положено $m_x = 0$.

естественно, требует объяснения. Происхождение (2.16) принято относить на счет *предельных теорем* о суммах независимых случайных величин. Об этом речь будет идти в следующих главах, но есть и другие причины, которые представляются не менее важными.

Как это часто бывает, многое становится ясным при помещении задачи в более широкий контекст.

Вместо случайной величины рассмотрим случайный вектор

$$x = \{x_1, \ldots, x_n\}$$

с *независимыми* координатами x_i и плотностью распределения $\rho(x)$, не зависящей от направления[12] x.

Этих «необременительных» предположений достаточно, чтобы гарантировать нормальное распределение всех x_i. Обоснование несложно. Независимость координат означает

$$\rho(x) = \rho_1(x_1) \ldots \rho_n(x_n), \tag{2.17}$$

а независимость $\rho(x)$ от направления x — постоянство плотности $\rho(x)$, равно как и ее логарифма

$$\ln \rho(x) = \ln \rho_1(x) + \ldots + \ln \rho_n(x) \tag{2.18}$$

на сферах $x_1^2 + \ldots + x_n^2 = \text{const}$.

Другими словами, функции (2.18) и $x^2 = x_1^2 + \ldots + x_n^2$ имеют одни и те же поверхности уровня, а это возможно, лишь когда их нормали (градиенты) коллинеарны (одинаково или противоположно направлены), т. е.

$$\nabla \ln \rho(x) = \lambda \nabla x^2,$$

что дает n равенств

$$\frac{\rho_i'(x_i)}{\rho_i(x)} + 2\lambda x_i = 0,$$

интегрирование которых приводит к $\ln \rho_i(x_i) = -\lambda x_i^2 + \text{const}$, т. е.

$$\rho_i(x_i) = \mu_i e^{-\lambda x_i^2}.$$

Константы определяются нормировкой и заданием, например, второго момента

$$\int_{-\infty}^{\infty} \mu_i e^{-\lambda x_i^2}\, dx = 1, \qquad \int_{-\infty}^{\infty} x_i^2 \mu_i e^{-\lambda x_i^2}\, dx_i = \sigma^2.$$

Окончательно

$$\rho(x) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{x_1^2 + \ldots + x_n^2}{\sqrt{2\pi\sigma^2}} \right\}. \tag{2.19}$$

[12] Можно иметь в виду стрельбу по плоской мишени с вертикальным отклонением x_1 и горизонтальным — x_2.

Под нормальным распределением случайного вектора в общем случае вместо (2.19) подразумевают плотность (2.17) с нормальными плотностями

$$\rho_i(x_i) = \frac{1}{\sigma_{x_i}\sqrt{2\pi}} \exp\left\{-\frac{(x_i - m_{x_i})^2}{2\sigma_{x_i}^2}\right\},$$

т. е.

$$\rho_X(x) = \frac{1}{\sqrt{(2\pi)^n \det K_x}} \exp\left\{-\frac{1}{2}(x - m_x)^T K_x^{-1}(x - m_x)\right\},$$

где K_x — ковариационная матрица, которая в данном случае диагональна, с элементами $\sigma_{x_i}^2$ на диагонали.

Если $\rho_X(x)$ — плотность случайного вектора $x = \{x_1, \ldots, x_n\}$, то

$$\rho_Y(y) = \frac{1}{|\det A|}\rho_X(A^{-1}y)$$

представляет собой плотность случайного вектора $Y = AX$, где A — невырожденная матрица.

Линейное преобразование $Y = AX$ нормально распределенного вектора X приводит к плотности

$$\rho_Y(y) = \gamma \exp\left\{-\frac{1}{2}(y - a)^T K_y^{-1}(y - a)\right\},$$

с коэффициентом γ, определяемым нормировкой плотности, а

$$K_y^{-1} = (A^{-1})^T K_x^{-1} A^{-1}.$$

В соответствии с этим вектор Z считается нормально распределенным, если его плотность равна

$$\rho_Z(z) = \frac{1}{\sqrt{(2\pi)^n \det K_z}} \exp\left\{-\frac{1}{2}(z - m_z)^T K_z^{-1}(z - m_z)\right\}, \qquad (2.20)$$

где K_z, как уже ясно, ковариационная матрица [13].

При данном способе изложения понятно, что *многомерный* нормальный закон распределения (2.20) при обратном линейном преобразовании снова возвращается к форме с нормально распределенными *независимыми* координатами.

[13] Разумеется, в (2.20) предполагается невырожденность K_z. Знак модуля с определителя K_z снят, поскольку $\det K_z > 0$ в силу положительной определенности K_z, см. раздел 1.8.

Философски настроенной части населения больше нравится интерпретация нормального закона как распределения, максимизирующего энтропию (глава 8). Точнее говоря, $\mathcal{N}(m, \sigma^2)$ есть решение оптимизационной задачи:

$$H = \int_{-\infty}^{\infty} \rho(x) \ln \rho(x)\, dx \to \max$$

при ограничениях

$$\int_{-\infty}^{\infty} \rho(x)\, dx = 1, \qquad \int_{-\infty}^{\infty} x\rho(x)\, dx = m, \qquad \int_{-\infty}^{\infty} x^2\rho(x)\, dx = \sigma^2 + m^2.$$

Складывая H с ограничениями, умноженными на множители Лагранжа λ, μ, ν, и варьируя $\rho(x)$, получаем нулевую вариацию Лагранжиана:

$$\int_{-\infty}^{\infty} \{(1 + \ln \rho(x)) + \lambda x^2 + \mu x + \nu\} \Delta\rho(x)\, dx = 0,$$

откуда, в силу произвольности $\Delta\rho(x)$, следует

$$1 + \ln \rho(x) + \lambda x^2 + \mu x + \nu = 0 \quad \Rightarrow \quad \rho(x) = e^{-\lambda x^2 - \mu x - \nu - 1}.$$

После согласования значений параметров с ограничениями задачи — получается плотность, соответствующая $\mathcal{N}(m, \sigma^2)$.

2.8. Пуассоновские потоки

Последовательность событий, происходящих в случайные моменты времени, называют *потоком событий*. Это один из мощных пластов вероятностных задач. Телефонные вызовы, аварии, обращения к оперативной памяти, заявки, посетители — список примеров практически неисчерпаем.

Рассмотрим поток событий, обладающий следующими свойствами:

- количества событий, поступающие на непересекающихся интервалах времени, независимы как случайные величины;
- вероятность поступления одного события за малый промежуток Δt зависит только от длины промежутка и равна $\lambda\Delta t + o(\Delta t)$, где $\lambda > 0$, $o(\Delta t)$ — бесконечно малая от Δt;
- вероятность поступления более одного события за время Δt есть $o(\Delta t)$.

◀ Разобьем интервал $(0, t)$ на n равных частей $\Delta_1, \ldots, \Delta_n$, и пусть $X_k(\Delta_k)$ обозначает число поступивших событий на промежутке Δ_k,

$$\mathbf{P}\{X_k(\Delta_k) = 1\} = \lambda t\left(\frac{1}{n}\right) + o\left(\frac{1}{n}\right).$$

В соответствии со сделанными предположениями производящая функция $\Pi_n(z)$ последовательности $p_j = \mathbf{P}\{X_k(\Delta_k) = j\}$ не зависит от k и равна

$$\Pi_n(z) = \left[1 - \lambda t\left(\frac{1}{n}\right)\right] + \lambda t\left(\frac{1}{n}\right)z + o\left(\frac{1}{n}\right).$$

Производящая функция суммы $X_1(\Delta_1) + \ldots + X_n(\Delta_n)$ вычисляется как произведение

$$\Pi(z) = [\Pi_n(z)]^n = \left[1 + \lambda t\left(\frac{1}{n}\right)(z-1) + o\left(\frac{1}{n}\right)\right]^n.$$

В пределе при $n \to \infty$ получается производящая функция числа событий $X(t)$, поступивших на интервале $(0, t)$:

$$\Pi(z) = \lim_{n\to\infty}\left[1 + \lambda t\left(\frac{1}{n}\right)(z-1)\right]^n = e^{\lambda t(z-1)}.$$

Соответствующее распределение вероятностей

$$\mathbf{P}\{X(t) = j\} = \frac{(\lambda t)^j}{j!}e^{-\lambda t} \qquad (j = 0, 1, 2, \ldots)$$

оказывается пуассоновским. Постоянная λ определяет среднюю интенсивность поступления событий, $\lambda = \dfrac{\mathbf{E}\,X(t)}{t}$. ▶

Проведенное рассуждение имеет пробел. Необходимо, вообще говоря, обосновать, что из сходимости производящих функций вытекает сходимость распределений вероятности. В данном случае это достаточно просто, но в принципе такого сорта вопросы постоянно возникают в **ТВ** — см. главу 4.

Рассмотренная задача легко обобщается на случай интенсивности λ, зависящей от времени. Результирующее распределение остается пуассоновским с учетом небольшой поправки:

$$\mathbf{P}\{X(t) = j\} = \frac{(\mu)^j}{j!}e^{-\mu}, \qquad \mu = \int_0^t \lambda(\tau)\,d\tau. \tag{2.21}$$

Временна́я интерпретация t, разумеется, необязательна. Речь может идти о распределении точек на любой числовой оси. А если вдуматься, то размерность t тоже не играет роли. *Распределение Пуассона возникает и в случае распределения точек в пространстве* при тех же исходных предположениях, в которых под Δt надо лишь понимать малые объемы. В итоге случайное число точек в области Ω

снова подчиняется распределению (2.21), с той разницей, что μ определяется как

$$\mu = \int_{\Omega} \lambda(\tau)\, d\tau.$$

Опыт общения со студентами показывает, что закон Пуассона часто остается «вещью в себе», не находя путей к подсознанию. Положение легко выправляется размышлением над задачами.

Допустим, случайная величина ξ, равномерно распределенная на $(0, T)$, реализуется n раз, что приводит к появлению на промежутке n точек. Сколько точек попадает в область

$$\Omega \in (0, T)?$$

Конечно, это в чистом виде схема Бернулли с вероятностью попадания отдельной точки в Ω, равной $p = l/T$, где l длина (мера) Ω. Вероятность попадания k точек в Ω определяется биномиальным распределением $C_n^k p^k (1-p)^{n-k}$, и далее проторенным в разделе 2.1 путем можно переходить к распределению Пуассона.

Для подсознания важна интерпретация этого пути в исходных терминах. Интервал $(0, T)$ и количество «бросаний» n увеличиваются согласованно. Так, чтобы среднее число точек на единицу длины сохранялось. Вот, собственно, и вся специфика предельного перехода. Значения T и n увеличиваются в одинаковое число раз, и тогда предельное распределение числа «попаданий» в Ω оказывается пуассоновским.

Экспоненциальное распределение. При пуассоновском распределении вероятность отсутствия событий на $(0, t)$ равна $e^{-\lambda t}$. Поэтому, если с. в. $t_1 > 0$ — это время наступления первого события, то $\mathbf{P}\{t_1 > t\} = e^{-\lambda t}$, а значит,

$$\mathbf{P}\{t_1 < t\} = 1 - e^{-\lambda t}. \tag{2.22}$$

Дифференцирование (2.22) по t дает плотность *экспоненциального закона*

$$\rho(t) = \lambda e^{-\lambda t}, \quad t \geqslant 0.$$

Из сказанного очевидно, что экспоненциальный закон представляет собой непрерывный аналог геометрического распределе-

ния — времени наступления первого успеха (события, поступления первой заявки, рекламации и т. п.).

Показательное распределение случайного времени ожидания возникает в ситуации, когда ожидание в течение времени s не влияет на то, сколько еще придется ждать, т. е.

$$\mathbf{P}\{\tau > s + t | \tau > s\} = \mathbf{P}\{\tau > t\},$$

что и приводит к $\mathbf{P}\{\tau > t\} = e^{-\lambda t}$, $t \geqslant 0$.

2.9. Статистики размещений

Как путь в большой спорт пролегает через подтягивание на турнике, так и в теории вероятностей есть простые модели того же назначения. Одна из них — размещение шаров по ячейкам. Шары, как и ячейки, могут быть неразличимы либо пронумерованы. Шаров r, ячеек n. Для создания атмосферы важности говорят о размещении элементарных частиц по энергетическим уровням. Не менее значимо иногда распределение карт между игроками, людей по месту работы, аварий по дням недели и т. п.

Если шары и ячейки различимы, то в n ячейках $r_1, \ldots, r_n$ шаров могут быть размещены числом способов $\dfrac{r!}{r_1! \ldots r_n!}$. Если все такие способы равновероятны, а их всего n^r, то соответствующее распределение имеет вид

$$P(r_1, \ldots, r_n) = \frac{r!}{r_1! \ldots r_n!}\, n^{-r}$$

и называется *статистикой Максвелла—Больцмана.*

Если шары (частицы) неразличимы, то число $B(r, n)$ всевозможных распределений равно числу целых решений $r_1, \ldots, r_n$ уравнения $r_1 + \ldots + r_n = r$. Это самостоятельная комбинаторная задача.

◀ Воспользуемся, для тренировки, методом производящих функций [14]. Очевидно, функция

$$\Pi(z) = \left(\sum_{r_1=0}^{\infty} z^{r_1}\right) \ldots \left(\sum_{r_n=0}^{\infty} z^{r_n}\right) = (1 - z)^{-n}, \qquad |z| < 1,$$

[14] Задача совсем просто решается переформулировкой. Выстраиваем шары в ряд, и делим их на n групп $n - 1$ запятыми. Число различимых способов: C_{r+n-1}^{r}.

при разложении в ряд порождает нужные коэффициенты,

$$(1-z)^{-n} = \sum_{r=0}^{\infty} B(r,n) z^r.$$

В результате

$$B(r,n) = \frac{1}{r!}\Pi^{(r)}(0) = \frac{n(n+1)\dots(n+r-1)}{r!} = C^r_{r+n-1}. \quad \blacktriangleright$$

При условии равновероятности различных способов возникает распределение

$$P(r,n) = \frac{1}{B(r,n)} = \frac{r!(n-1)!}{(n+r-1)!},$$

называемое *статистикой Бозе—Эйнштейна.*

При дополнительном запрете «ячейка не может содержать более одного шара» получается *статистика Ферми—Дирака*:

$$P(r,n) = \frac{1}{C^r_n} = \frac{r!(n-r)!}{n!}, \qquad r \leqslant n.$$

Все очень просто, но в этом и заключается секрет «подтягивания на турнике». Рутинная возня с простыми моделями дает навык обращения с различимыми и неразличимыми вариантами. А это как раз граница между правильными и неправильными решениями.

2.10. Распределение простых чисел

Вероятностный стиль рассуждений эффективно работает и на «чужих территориях», где все детерминированно. Показательна в этом отношении *задача о распределении простых чисел.*

Количество простых чисел [15], не превосходящих x, — принято обозначать через $\pi(x)$. С указанием всех простых чисел легко (идеологически) справляется *решето Эратосфена*, рецепт которого очень прост. Из записи всех натуральных чисел вычеркивается 1 — первое невычеркнутое число 2 — простое. Далее зачеркиваются числа, делящиеся на 2, число 3 — первое невычеркнутое — простое. И так далее.

[15] Не имеющих, по определению, делителей кроме 1 и самого себя.

При этом ясно, что в промежутке $[x, x + \Delta x]$ доля чисел, делящихся на простое p, равна $1/p$, а не делящихся — $(1 - 1/p)$. Доля же чисел в этом промежутке, не делящихся ни на одно простое число, равна

$$\rho(x) = \left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{3}\right) \ldots \left(1 - \frac{1}{p}\right), \tag{2.23}$$

причем ясно, что говорить имеет смысл о простых p меньших $\sqrt{x}$, и о $\Delta x \ll x$, но $\Delta x > \varepsilon x$ при некотором малом $\varepsilon > 0$.

Самих простых чисел на $[x, x + \Delta x]$ будет

$$\rho(x)\Delta x \approx \pi(x + \Delta x) - \pi(x),$$

т. е. $\rho(x)$ играет роль плотности, а формула (2.23) получается из «предположения о независимости» событий делимости любого натурального k на разные простые числа.

Дальнейшее опирается на манипуляции «асимптотического толка», несколько злоупотребляющие ссылками на здравый смысл.

Для больших p приближенно: $1 - 1/p = e^{-1/p}$. Поэтому

$$\ln \rho(x) = -\sum_k \frac{1}{p_k},$$

где p_k обозначает k-е простое число.

В промежутке $[x, x + \Delta x]$, в силу $\Delta x \ll x$, можно считать $p_k \sim x$, и сумма по этому промежутку

$$\sum \frac{1}{p_k} \sim \frac{1}{x}\rho(x)\Delta x,$$

откуда

$$\ln \rho(x) = -\sum_k \frac{1}{p_k} \sim -\int_1^x \frac{\rho(u)}{u}\, du,$$

что после дифференцирования по x приводит к уравнению

$$\frac{\rho'(x)}{\rho(x)} = -\frac{\rho(x)}{x} \quad \Rightarrow \quad \frac{d\rho}{\rho^2} = -\frac{dx}{x},$$

решение которого $\rho(x) = 1/(C + \ln x)$ при больших x переходит в

$$\boxed{\rho(x) = \frac{1}{\ln x}.}$$

Что касается $\pi(x)$, то

$$\pi(x) = \int_2^x \frac{du}{\ln u} = \frac{x}{\ln x}\left\{1 + \frac{1}{\ln x} + \ldots + \frac{r!}{\ln^r x} + O\left(\frac{1}{\ln^{r+1} x}\right)\right\}. \qquad (2.24)$$

Для примера, точное значение $\pi(4000) = 550$. Первые три члена разложения (2.24) дают приближение $\pi(4000) \approx 554$.

2.11. Задачи и дополнения

- Будем считать в данном пункте, что речь идет о случайных векторах с нулевыми матожиданиями. Любой случайный вектор X линейным преобразованием $Y = AX$ приводится к вектору Y с *некоррелированными* координатами [16]. Действительно, матожидание матричного равенства

$$YY^T = AXX^TA^T$$

дает $K_y = AK_xA^T$. Неотрицательно определенная ковариационная матрица K_x всегда может быть приведена [17] ортогональным преобразованием A к диагональному виду K_y.

- Плотность *распределения Коши*

$$\rho(x) = \frac{1}{\pi(1+x^2)} \qquad (2.25)$$

имеет с. в. $X = \zeta_1/\zeta_2$, где независимые с. в. ζ_1 и ζ_2 распределены нормально по закону $\mathcal{N}(0, 1)$. Такая же плотность распределения у $\operatorname{tg}\theta$ при условии равномерного распределения θ на $[-\pi/2, \pi/2]$.
Распределение Коши имеет дурную славу, поскольку обычно извлекается на свет, когда надо продемонстрировать существование «плохих» законов, не имеющих моментов.

- Если все с. в. $X_1, \ldots, X_n$ независимы и имеют одинаковое распределение (2.25), то

$$Y_n = \frac{X_1 + \ldots + X_n}{n}$$

распределена по тому же закону. (**?**)

- Если $f(x_1, \ldots, x_n)$ — совместная плотность вектора $X = \{X_1, \ldots, X_n\}$, то сумма $S = \sum_1^n X_k$ имеет плотность

$$\rho(s) = \int \ldots \int f\Big(s - \sum_2^n x_k, x_2, \ldots, x_n\Big)\, dx_2 \ldots dx_n.$$

---

[16] А в случае нормально распределенного X — к вектору Y с независимыми координатами.

[17] См. [5, т. 3].

- Проекция радиус-вектора X, равномерно распределенного на окружности радиуса r, имеет функцию распределения

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{x}{r}, \quad x \in (-r, r),$$

при естественном условии $F(x \leqslant -r) = 0$ и $F(x \geqslant r) = 1$. (?)
Соответствующая плотность:

$$\rho(x) = F'(x) = \frac{1}{\pi\sqrt{r^2 - x^2}}, \quad x \in (-r, r).$$

Следствие приведенных формул: при равномерном вращении коленчатого вала поршни двигателя внутреннего сгорания бóльшую часть времени проводят в крайних положениях.
Аналогичное явление наблюдается при игре в «орлянку».

- Случайные величины X, Y независимы и равномерно распределены на $[-1/2, 1/2]$ (каждая). Плотность распределения произведения $Z = XY$ равна

$$\rho(z) = -2 \ln 4|z|, \quad |z| \leqslant \frac{1}{4}. \quad (?)$$

- У нормально распределенного вектора $\{X, Y\}$ с плотностью

$$\rho(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

полярные координаты R, Φ в представлении

$$X = R\cos\Phi, \quad Y = R\sin\Phi$$

распределены: Φ равномерно на $[0, 2\pi]$, а R по *закону Рэлея*

$$\rho(r) = \frac{r}{\sigma^2} e^{-r^2/(2\sigma^2)}.$$

- Пусть независимые с. в. $X_1, \ldots, X_n$ имеют показательные распределения с параметрами $\lambda_1, \ldots, \lambda_n$. Тогда

$$X = \min\{X_1, \ldots, X_n\}$$

имеет показательное распределение с параметром $\lambda = \lambda_1 + \ldots + \lambda_n$.

- Задач на определение тех или иных вероятностных распределений имеется великое множество, и за ними далеко ходить не надо. В любой стандартной модели «шаг влево или вправо» — и возникают неясности. При бросании монеты (случайном блуждании) — биномиальное распределение, казалось бы, исчерпывает проблематику. Но это далеко не так. Вопрос о первом успехе — и появляется геометрическое распределение. Механика смены лидерства (перехода блуждающей частицы слева направо или наоборот) — и дорога уводит к *закону арксинуса*. Вопрос о локальных экстремумах либо попадании траектории выигрыша на некоторую кривую — и опять новые законы распределения.

Самые простые вопросы порождают иногда очень сложные задачи. *Модель Изинга*, например. Если молекулы двух типов «1», «2» располагаются в шеренгу (одномерная модель), то энергия цепочки равна

$$H = \sum_{i,j=1}^{2} n_{ij} H_{ij},$$

где H_{ij} — энергия взаимодействия соседних молекул, в случае, когда за молекулой типа «i» следует молекула типа «j».
Равновероятное расположение молекул порождает легко определяемое распределение H. (?) Но уже двумерная (а тем более — трехмерная) модель — не поддается исчерпывающему анализу [18].

- Источником для упражнений может служить любая задача, в том числе классическая — что лучше всего. Вот одна из таких задач.
Имеются две одинаковые колоды, в каждой N карт, нумеруемых в порядке их случайного расположения. Если событие A_k обозначает совпадение карт первой и второй колоды, расположенных на k-м месте, то, очевидно,

$$\mathbf{P}\{A_k\} = \frac{1}{N!}, \quad \mathbf{P}\{A_i A_j\} = \frac{(N-1)!}{N!} = \frac{1}{N},$$
$$\mathbf{P}\{A_i A_j A_k\} = \frac{(N-2)!}{N!} = \frac{1}{N(N-1)}, \quad \ldots,$$

и применение формулы (1.3) сразу дает

$$p_1 = \mathbf{P}\left\{\sum_{k}^{N} A_k\right\} = 1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^{N-1}\frac{1}{N!},$$

т. е. вероятность, что совпадет хотя бы одна карта, равна $p_1 \approx e^{-1}$.
Остается задача вычисления полного распределения $p_1, \ldots, p_N$. (?) Среднее число совпадений,

$$\sum_{k} k p_k = 1,$$

легко определяется окольным рассуждением. (?)

[18] Задача не решена, несмотря на ее значимость для кристаллографии и ферромагнетизма.

Глава 3

Законы больших чисел

Идея рождения порядка из хаоса материализуется в нашем мире различными способами. Один из наглядных вариантов — закон больших чисел, который, так или иначе, говорит о стабилизации средних значений.

3.1. Простейшие варианты

Если случайные величины X_i имеют одно и то же математическое ожидание μ, то

$$\frac{S_n}{n} = \frac{X_1 + \ldots + X_n}{n}$$

имеет то же самое матожидание μ, и с ростом n при естественных предположениях «становится все менее случайной величиной». Различные варианты уточнения этого утверждения называют *законом больших чисел.*

3.1.1. *Пусть некоррелированные случайные величины X_i имеют одно и то же матожидание μ и одну и ту же дисперсию σ^2. Тогда среднеквадратичное отклонение S_n/n от матожидания стремится к нулю. Точнее,*

$$\boxed{\mathbf{D}\left\{\frac{S_n}{n}\right\} = \frac{\sigma^2}{n} \to 0 \quad \textit{при} \quad n \to \infty,}$$

причем $S_n/\sqrt{n}$ растет в среднем пропорционально $\mu\sqrt{n}$, имея постоянную дисперсию σ^2.

◀ В силу некоррелированности,

$$\mathbf{E}\,(X_i - \mu)(X_j - \mu) = 0 \quad \text{при} \quad i \neq j.$$

Поэтому

$$\mathbf{D}\left\{\frac{S_n}{n}\right\} = \mathbf{E}\left\{\frac{\sum_i (X_i - \mu)}{n}\right\}^2 = \mathbf{E}\left\{\frac{\sum_i (X_i - \mu)^2}{n^2}\right\} = \frac{\sigma^2}{n}.$$

Аналогично рассматривается $S_n/\sqrt{n}$. ▶

В комбинации с утверждением 3.1.1 неравенство Чебышева (1.16) приводит к другому варианту закона больших чисел.

3.1.2. *Пусть некоррелированные случайные величины X_i имеют одно и то же математическое ожидание μ и одну и ту же дисперсию σ^2. Тогда при любом $\varepsilon > 0$*

$$\boxed{\mathbf{P}\left\{\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| > \varepsilon\right\} \leqslant \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \textit{при} \quad n \to \infty.}$$

Закон больших чисел сводит концы с концами. Частотная трактовка вероятности (1.1) приобретает законную силу, что устанавливает связь между абстрактными моделями и статистическими экспериментами. Если в случайной «01»-последовательности единица ($X_i = 1$) появляется с вероятностью 1/2, то вероятность отклонения среднего S_n/n от математического ожидания 1/2 более чем на 0,1 — не превосходит $25/n$, поскольку в данном случае

$$\sigma^2 = \left(1 - \frac{1}{2}\right)^2 \cdot \frac{1}{2} + \left(0 - \frac{1}{2}\right)^2 \cdot \frac{1}{2} = \frac{1}{4}.$$

Предположения в 3.1.1, 3.1.2 о том, что величины X_i имеют одинаковые математические ожидания и дисперсии, разумеется, необязательны. Тот же метод доказательства работает и в более общих ситуациях. Например, при некоррелированности $X_1, \ldots, X_n$ и

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 = 0,$$

где σ_i^2 — дисперсия X_i, при любом $\varepsilon > 0$ имеет место

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{X_1 + \ldots + X_n}{n} - \frac{\mu_1 + \ldots + \mu_n}{n}\right| > \varepsilon\right\} = 0,$$

где μ_i — математическое ожидание X_i.

Рассмотренные варианты стабилизации среднего обычно характеризуются как *слабый закон больших чисел*, но он при осознании производит довольно сильное психологическое впечатление. Вплоть до убеждения в одушевленности механизма обращения случайных цепочек в целесообразные явления. Это, конечно, вредит пониманию существа дела и плодит иллюзии.

3.2. Усиленный закон больших чисел

Слабый закон больших чисел дает оценки вероятности отклонений среднестатистических сумм от матожидания и гарантирует стремление этих вероятностей к нулю. Усиленный закон больших чисел дает больше, гарантирует равенство предела среднестатистической суммы матожиданию — с вероятностью 1. На первых порах знакомства с **ТВ** разница обычно не чувствуется, но она довольно существенна, что выявляется при рассмотрении различных видов вероятностной сходимости — см. главу 4.

Основой для анализа событий, происходящих «почти наверное», является следующий простой факт.

3.2.1 Лемма Бореля—Кантелли.

(I) *В любой последовательности событий* $A_1, A_2, \ldots$ *— при условии* $\sum_{k=1}^{\infty} \mathbf{P}(A_k) < \infty$ *— с вероятностью* 1 *происходит лишь конечное число событий* A_n.

(II) *В любой последовательности* $A_1, A_2, \ldots$ ***независимых*** *событий — при условии* $\sum_{k=1}^{\infty} \mathbf{P}(A_k) = \infty$ *— с вероятностью* 1 *происходит бесконечное число событий* A_n.

◀ (i). Наступление бесконечного числа $A_1, A_2, \ldots$ есть событие

$$A = \bigcap_n \Big(\bigcup_{k \geqslant n} A_k \Big).$$

А поскольку

$$\mathbf{P}(A) \leqslant \mathbf{P}\Big(\bigcup_{k \geqslant n} A_k \Big) \leqslant \sum_{k \geqslant n} \mathbf{P}(A_k) \to 0 \quad \text{при} \quad n \to \infty,$$

то $\mathbf{P}(A) = 0$, что влечет за собой $\mathbf{P}(\bar{A}) = 1$. ▶

◀ Для доказательства (ii) достаточно проверить условие

$$\mathbf{P}\Big(\bigcup_{k\geqslant n} A_k\Big) = 1 \quad \text{для любого} \quad n, \tag{3.1}$$

поскольку пересечение множеств меры 1 должно иметь ту же полную меру 1.

В силу $\sum_{k=1}^{\infty} \mathbf{P}(A_k) = \infty$ и независимости A_n, а значит и $\bar{A}_n = \Omega\backslash A_n$, для любого $N \geqslant k$:

$$1 - \mathbf{P}\Big(\bigcup_{k\geqslant n} A_k\Big) \leqslant 1 - \mathbf{P}\Big(\bigcup_{n\leqslant k\leqslant N} A_k\Big) = \mathbf{P}\Big(\bigcup_{n\leqslant k\leqslant N} \bar{A}_k\Big) = \prod_{n\leqslant k\leqslant N} [1 - \mathbf{P}(A_k)] \leqslant$$

$$\leqslant \exp\Big\{-\sum_{n\leqslant k\leqslant N} \mathbf{P}(A_k)\Big\} \to 0$$

при $N \to \infty$, что влечет за собой (3.1). ▶

Приведем теперь один из простейших вариантов усиленного закона больших чисел.

*Пусть некоррелированные случайные величины X_i имеют нулевое математическое ожидание и **конечный четвертый момент**. Тогда*

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} = 0\right) = 1. \tag{3.2}$$

◀ Число ненулевых слагаемых в $\mathbf{E}(X_1 + \ldots + X_n)^4$ — после раскрытия скобок, — в силу некоррелированности, пропорционально n^2. А ограниченность четвертых моментов гарантирует при этом существование константы C такой, что $\mathbf{E}(X_1 + \ldots + X_n)^4 \leqslant Cn^2$. Поэтому (см. раздел 1.9)

$$\mathbf{P}(|X_1 + \ldots + X_n| \geqslant \varepsilon n) \leqslant \frac{Cn^2}{(\varepsilon n)^4} = \frac{C}{(\varepsilon n)^2}.$$

В силу $\sum_{n}^{\infty} \frac{C}{(\varepsilon n)^2} < \infty$ лемма 3.2.1 гарантирует конечность числа событий

$$\frac{|X_1 + \ldots + X_n|}{n} > \varepsilon,$$

откуда в конечном итоге следует (3.2). ▶

Смысл предположения ограниченности четвертых моментов достаточно прозрачен. Иначе, при том же методе доказательства, не удалось бы установить сходимость ряда

$$\sum_{n}^{\infty} \mathbf{P}(|X_1 + \ldots + X_n| \geqslant \varepsilon n)$$

и, соответственно, воспользоваться леммой Бореля—Кантелли.

Но сам по себе рассмотренный вариант усиленного закона больших чисел довольно слаб. Более тонкие рассуждения дают те же выводы в менее ограничительных предположениях.

3.2.2 Теорема [1]. *Пусть независимые величины X_n имеют математические ожидания μ_n и дисперсии σ_n^2. При условии* $\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty$ *имеет место*

$$\frac{X_1 + \ldots + X_n}{n} - \frac{\mu_1 + \ldots + \mu_n}{n} \to 0 \qquad (n \to \infty)$$

с вероятностью единица.

Одно из классических применений усиленного закона больших чисел указано Борелем. Число $x \in [0, 1]$ называется *нормальным*, если при его записи в любой d-ичной системе счисления частота появления каждой цифры равна $1/d$. Доказательство нормальности почти всех $x \in [0, 1]$ (за исключением множества меры 0) см. например, в [11, 15].

3.3. Нелинейный закон больших чисел

Понятно, что вместо стабилизации среднего предпочтительнее было бы говорить об условиях стабилизации нелинейных функций

$$y = f_n(x_1, \ldots, x_n),$$

становящихся при больших n почти константами. Формальная постановка вопроса могла бы опираться на следующее определение. Последовательность функций $f_n(x)$ *асимптотически постоянна*, если существует такая числовая последовательность μ_n, что

$$\mathbf{P}\{|f_n(x) - \mu_n| > \varepsilon\} \to 0 \quad \text{при} \quad n \to \infty \tag{3.3}$$

для любого наперед заданного $\varepsilon > 0$ [2]. Либо, в более жестком варианте, можно потребовать $\mathbf{D}\{f_n(x)\} \to 0$ при $n \to \infty$.

Классическая теория вероятностей имеет хорошие ответы на вопрос о справедливости (3.3) в случае $f_n(x) = \sum x_i$. Вот простейшая формулировка с некоторым отступлением от стандарта.

[1] См., например, [18, 21].

[2] Под $\mathbf{P}\{\cdot\}$ подразумевается некоторая заданная мера.

Пусть x_i — независимые с. в. с одинаковыми матожиданиями μ_x и дисперсиями $D_x = \sigma_x^2$. Тогда дисперсия линейной функции

$$y = c(n) \cdot x = c_1 x_1 + \ldots + c_n x_n$$

равна

$$D_y = c^2(n) D_x = (c_1^2 + \ldots + c_n^2) D_x,$$

и в результате $D_y \to 0$ при условии $\|c(n)\| \to 0$ $(n \to \infty)$.

При изучении нелинейных зависимостей $y = f_n(x)$ под тем же углом зрения естественно взять за основу аналогичные ограничения на градиент

$$\nabla f_n(x) = \left\{ \frac{\partial f_n}{\partial x_1}, \ldots, \frac{\partial f_n}{\partial x_n} \right\}.$$

Заметим, что для гладких функций условие $\|\nabla f_n(x)\| \leqslant \gamma_n$ эквивалентно в R^n липшицевости $f_n(x)$ с константой γ_n,

$$\|\nabla f_n(x) - \nabla f_n(y)\| \leqslant \gamma_n \|x - y\|. \tag{3.4}$$

Если же условию (3.4) удовлетворяет негладкая функция, то у нее существует сколь угодно точная гладкая аппроксимация с модулем градиента $\leqslant \gamma_n$.

Рассмотрим теперь для наглядности простейший случай с. в. X_i, равномерно распределенных на $[0, 1]$. Другими словами, равномерное распределение на кубе

$$C_n = [0, 1] \times \ldots \times [0, 1].$$

О вероятностной точке зрения, собственно, можно забыть[3]. Задана последовательность функций $f_n(x)$, и мы интересуемся условиями, при которых отклонение $f_n(x)$ от среднего значения стремится к нулю с ростом n.

Естественный ориентир задает линейный случай. Но хватит ли ограниченности градиента $\|\nabla f_n(x)\|$ для $\mathbf{D}\{f_n\} < \infty$ в общей ситуации? Ведь разброс значений $f_n(x)$ — разность между минимумом и максимумом — может расти пропорционально диаметру куба C_n, т. е. $\sqrt{n}$.

Если ответ положителен (а он *положителен*), то можно ли перейти к какой либо другой мере на C_n, не потеряв желаемых

[3] В этом, грубо говоря, и заключается идея Колмогорова изучать теорию вероятностей как часть теории меры.

выводов? Конечно, к произвольной мере перейти нельзя, иначе, сосредоточив ее на концах большой диагонали куба, получим $\mathbf{D}\{f_n\} \sim n$. Но достаточно ли, скажем, независимости X_i? Какие плотности дают максимум $\mathbf{D}\{f_n\}$? Как от куба перейти к рассмотрению всего пространства? Вот примерный круг вопросов, которые здесь возникают.

Откладывая доказательство, сформулируем следующий результат.

3.3.1 Теорема. *Пусть независимые с. в. X_i распределены на $[0, 1]$ с плотностями $\rho_i(x_i)$, причем все $\rho_i(x_i) > \varepsilon > 0$, а последовательность функций $f_n(x_1, \ldots, x_n)$ удовлетворяет неравенствам*

$$\|\nabla f_n(x)\| \leqslant \gamma_n, \quad x \in C_n.$$

Тогда при $\gamma_n < \gamma < \infty$ дисперсия $\mathbf{D}\{f_n\}$ ограничена некоторой константой, не зависящей от n. Если же γ_n стремится к нулю с ростом n, то

$$\mathbf{D}\{f_n\} \to 0 \quad \textit{при} \quad n \to \infty,$$

т. е. последовательность функций $f_n(x)$ асимптотически постоянна на C_n.

Рекламный вариант теоремы мог бы звучать так: *все липшицевы функции большого числа переменных — константы.*

3.4. Оценки дисперсии

В формулировке дальнейших результатов принимает участие нестандартная для теории вероятностей функция

$$\rho^*(x) = \mu(\infty) \int_{-\infty}^{x} \rho(t)\, dt - \mu(x), \qquad \mu(x) = \int_{-\infty}^{x} t\rho(t)\, dt, \tag{3.5}$$

которую назовем *сопряженной плотностью*[4].

Эффективный способ оценки дисперсии дает следующее утверждение.

[4] Сопряженная плотность может быть ненормирована.

3.4.1 Лемма. *Пусть с. в. X_i распределены независимо с плотностями $\rho_i(x_i)$, каждая из которых имеет сопряженную $\rho_i^*(x_i)$. Тогда для любой непрерывно дифференцируемой функции $f(x_1, \ldots, x_n)$ справедливо неравенство*

$$\mathbf{D}(f) \leqslant \int_{R^n} \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \rho_i^*(x_i) \prod_{j \neq i} \rho_i(x_i)\, dx_1 \ldots dx_n, \tag{3.6}$$

при условии существования фигурирующего в (3.6) *интеграла как повторного.*

Доказательство приводится в следующем разделе.

Требование существования сопряженной плотности равносильно ограничению на порядок убывания обычных плотностей $\rho(x)$ на бесконечности, что может выражаться в терминах существования моментов. Легко убедиться, что при переходе от $\rho(x)$ к $\rho^*(x)$ порядок убывания «ухудшается на единицу». Если, например, $\rho(x) = o(|x|^{-k})$, то $\rho^*(x) = o(|x|^{-k+1})$ при $|x| \to \infty$.

Для $\rho(x) = \frac{1}{2}e^{-|x|}$ сопряженная плотность

$$\rho^*(x) = \frac{1}{2}(1 + |x|)e^{-|x|}.$$

Если же область определения $\rho(x)$ конечна, то сопряженная плотность существует всегда. Наиболее отчетливо ее роль выявляется в ситуациях типа

$$\rho(x) = p\delta(x) + (1 - p)\delta(x - 1),$$

где сопряженная плотность равномерна,

$$\rho^*(x) \equiv p(1 - p),$$

т. е. $\rho^*(x) > 0$ там, где $\rho(x) = 0$. Компенсирующий эффект при этом заключается в следующем. Если бы, скажем, в неравенстве (3.6) вместо ρ^* стояла исходная плотность ρ, то такое неравенство было бы заведомо ошибочно, поскольку $f(x)$ могла бы расти лишь там, где плотности $\rho_i(x) = 0$, и справа был бы 0 при $\mathbf{D}(f) > 0$. Сопряженная же плотность «следит» за поведением градиента $f(x)$ на тех участках, где исходная плотность обнуляется.

Когда X_i равномерно распределены на $[0, 1]$, сопряженные плотности $\rho^*(x) = \frac{1}{2}x(1 - x)$, и (3.6) переходит в

$$\mathbf{D}(f) \leqslant \int_{C^n} \sum_{i=1}^{n} x_i(1 - x_i) \left(\frac{\partial f}{\partial x_i}\right)^2 dx_1 \ldots dx_n, \tag{3.7}$$

что тем более влечет за собой

$$\mathbf{D}(f) \leqslant \frac{1}{8}\int_{C^n} [\nabla f(x)]^2\, dx_1 \ldots dx_n. \tag{3.8}$$

Константу в неравенстве (3.8) — которое естественно называть многомерным аналогом *неравенства Виртингера* [1] — можно уменьшить до π^{-2}, но в данном контексте это не представляет особого интереса.

Из (3.8) сразу следует, что при равномерном распределении x на C_n для $f_n(x)$ справедлив практически тот же результат, что и в линейном случае:

$$\mathbf{D}\{f_n\} \to 0, \quad \text{если} \quad \max_x \|\nabla f_n(x)\| \to 0 \quad \text{при} \quad n \to \infty.$$

Преимущества неравенства (3.7) выявляются на такой функции, как

$$f_n(x) = \max_i |x_i|, \quad x \in C_n.$$

Здесь $\|\nabla f_n(x)\| = 1$ почти везде, но (3.7) гарантирует (после аккуратных вычислений) $\mathbf{D}\{f_n\} \sim 1/n$.

Заметим, наконец, что из леммы 3.4.1 практически сразу вытекает теорема 3.3.1. Действительно, из (3.6) в предположениях теоремы следует оценка

$$\mathbf{D}\{f_n\} \leqslant \nu \max_{x \in C_n} [\nabla f(x)]^2,$$

где

$$\nu = \sup\left\{ \frac{\rho_i^*(x)}{\rho_i(x)} : x \in [0, 1],\ i = 1, \ldots, n \right\} < \infty,$$

что, собственно, и обеспечивает требуемые выводы.

3.5. Доказательство леммы 3.4.1

Докажем сначала (3.6) в одномерном случае. Очевидно,

$$\mathbf{D}\{f\} = \int_{-\infty}^{\infty} [f(x) - m_f]^2\, dP(x) = \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [f(x) - f(y)]^2\, dP(x)\, dP(y) =$$

$$= \int_{-\infty}^{\infty}\int_{y}^{\infty}\left[\int_{y}^{x} f'(t)\, dt\right]^2 dP(x)\, dP(y), \qquad \text{где} \quad dP(x) = \rho(x)\, dx.$$

Учитывая [5]

$$\left[\int_y^x f'(t)\,dt\right]^2 \leqslant (x-y)\int_y^x [f'(t)]^2\,dt, \tag{3.9}$$

получаем

$$\mathbf{D}\{f\} \leqslant \int_{-\infty}^{\infty}\int_y^{\infty}(x-y)\int_y^x [f'(t)]^2\,dt\,dP(x)\,dP(y). \tag{3.10}$$

Интегрирование в (3.10) идет по области, задаваемой неравенствами

$$-\infty < y \leqslant t \leqslant x < \infty.$$

Изменим порядок интегрирования на следующий. По x — от t до ∞, по y — от $-\infty$ до t, по t — от $-\infty$ до ∞. В результате имеем

$$\mathbf{D}\{f\} \leqslant \int_{-\infty}^{\infty}[f'(t)]^2\int_t^{\infty}\int_{-\infty}^{t}(x-y)\,dP(x)\,dP(y)\,dt = \int_{-\infty}^{\infty}[f'(t)]^2\rho^*(t)\,dt,$$

поскольку, как легко убедиться,

$$\int_t^{\infty}\int_{-\infty}^{t}(x-y)\,dP(x)\,dP(y) = \rho^*(t),$$

а изменение порядка в условиях леммы законно. Таким образом, (3.6) в одномерном случае установлено.

Далее действуем по индукции. Пусть (3.6) справедливо в размерности $n-1$. Введем обозначения

$$m_{n-1}(x_n) = \int_{R^{n-1}} f(x)\,dP_1(x_1)\ldots dP_{n-1}(x_{n-1}),$$

$$D_{n-1}(x_n) = \int_{R^{n-1}} [f(x) - m_{n-1}(x_n)]^2\,dP_1(x_1)\ldots dP_{n-1}(x_{n-1}).$$

Очевидно,

$$\mathbf{D}\{f\} = \int_{R^n}[f(x) - m_{n-1}(x_n) + m_{n-1}(x_n) - m_f]^2\,dP_1(x_1)\ldots dP_n(x_n) =$$

---

[5] Неравенство (3.9) получается, если в неравенстве Коши—Буняковского

$$\left[\int_y^x u(t)v(t)\,dt\right]^2 \leqslant \int_y^x u^2(t)\,dt\int_y^x v^2(t)\,dt$$

положить $u(t) = f'(t)$, $v(t) \equiv 1$.

$$= \int_{R^1} D_{n-1}(x_n)\, dP_n(x_n) + \int_{R^1} [m_{n-1}(x_n) - m_f]^2\, dP_n(x_n),$$

а из доказанного уже одномерного неравенства (3.6) следует

$$\int_{R^1} [m_{n-1}(x_n) - m_f]^2\, dP_n(x_n) \leqslant \int_{R^1} \left[\frac{dm_{n-1}(x_n)}{dx_n}\right]^2 \rho^*(x_n)\, dx_n.$$

Наконец

$$\left[\frac{dm_{n-1}(x_n)}{dx_n}\right]^2 = \left[\frac{d}{dx_n} \int_{R^{n-1}} f(x)\, dP_1(x_1) \ldots dP_{n-1}(x_{n-1})\right]^2 \leqslant$$

$$\leqslant \int_{R^{n-1}} \left(\frac{df}{dx_n}\right)^2 \rho_1(x_1) \ldots \rho_{n-1}(x_{n-1})\, dx_1 \ldots dx_{n-1}.$$

Проведенные выкладки в совокупности с индуктивным предположением обеспечивают справедливость (3.6) в размерности n. Лемма доказана.

3.6. Задачи и дополнения

- Несмотря на математическую тривиальность закона больших чисел, он довольно часто понимался превратно. Оправданием могут служить многочисленные «аномальные» эффекты в его окрестности (см. главу 4). Ограничимся упоминанием самых простых, но достаточно удивительных фактов.

При n бросаниях монеты серия из гербов длины $\log_2 n$ *наблюдается с вероятностью, стремящейся к* 1 *при* $n \to \infty$. (**?**)

На фоне обязательного присутствия длинных чистых серий (только гербы или только решетки) средняя длина чистой серии равна 2. *Для любой несимметричной монеты, выпадающей гербом с вероятностью* $p \in (0, 1)$, *математическое ожидание длины нечетных по числу бросаний серий равно*

$$\frac{p}{1-p} + \frac{1-p}{p},$$

а четных — равно 2 *независимо от* p. (**!?**)

Пусть $S_n = X_1 + \ldots + X_n$, где X_k принимают значения 1, 0 с вероятностями p_k, $1 - p_k$ (каждый раз бросается другая монета). При появления разброса вероятностей p_k относительно

$$p = \frac{p_1 + \ldots + p_n}{n}$$

дисперсия S_n уменьшается. (**!?**)

- Усиленный закон больших чисел имеет множество вариаций. Вот один из достаточно тонких результатов А. Н. Колмогорова, где не требуется существование вторых моментов.

  Пусть X_i независимые случайные величины с одинаковым распределением и математическим ожиданием μ. Тогда

$$\mathbf{P}\left(\lim_{n\to\infty}\frac{X_1+\ldots+X_n}{n}=\mu\right)=1.$$

  Если же математическое ожидание X_i не существует, то

$$\mathbf{P}\left(\overline{\lim_{n\to\infty}}\left|\frac{X_1+\ldots+X_n}{n}\right|=\infty\right)=1.$$

- Снять в теореме 3.3.1 ограничение $\rho_i(x) > \varepsilon > 0$ без каких либо компенсирующих предположений нельзя. При обнулении плотностей на множествах ненулевой меры дисперсия $f_n(x)$ может неограниченно возрастать при ограниченном по модулю градиенте. Рассмотрим, например, линейную функцию

$$\varphi_n(x)=x_1+(\sqrt{2}-1)x_2+\ldots+(\sqrt{n}-\sqrt{n-1})x_n,$$

  считая ее определенной лишь на вершинах куба C_n. Легко проверить, что на вершинах куба C_n

$$|\varphi_n(x)-\varphi_n(y)|\leqslant\|x-y\|. \tag{3.11}$$

  Продолжим $\varphi_n(x)$ с вершин куба на весь куб C_n с сохранением условия сжатия (3.11), что всегда возможно [7]. Пусть $f_n(x)$ — соответствующее продолжение. Поскольку $f_n(x)$ принимает на вершинах куба C_n те же значения, что и $\varphi_n(x)$, то в случае

$$\rho_i(x_i)=\frac{1}{2}\delta(x_i)+\frac{1}{2}\delta(x_i-1)$$

  дисперсии $f_n(x)$ и $\varphi_n(x)$ совпадают. Но дисперсия линейной функции легко считается. В результате

$$\mathbf{D}\{f_n\}=\mathbf{D}\{\varphi_n\}=\frac{1}{4}[\nabla\varphi_n(x)]^2\sim\ln n,$$

  т. е. $\mathbf{D}\{f_n\}\to\infty$ при $n\to\infty$.

- Пусть X_i распределены независимо на $[0,1]$ с произвольными плотностями, и пусть задана последовательность определенных на C_n липшицевых функций $f_n(x)$ с константами Липшица γ_n. Тогда

$$\mathbf{D}\{f_n\}\to 0,\quad \text{если}\quad \gamma_n=o\left(\frac{1}{\sqrt[4]{n}}\right).$$

  Если же $1/\sqrt[4]{n}=o(\gamma_n)$, то найдется такая последовательность $f_n(x)$, что $\mathbf{D}\{f_n\}\to\infty$.

- Изучение пограничных ситуаций типа рассмотренной в предыдущем пункте опирается на построение примеров последовательностей $f_n(x)$ с максимальными $\mathbf{D}\{f_n\}$. Довольно неожиданно, что функции $f_n(x)$, обеспечивающие максимум дисперсии, оказываются симметрическими [6]. Неожиданно — с позиций распространенной легенды, которая главным источником статистических закономерностей считает симметрию. Такой взгляд особенно упрочился после интересной публикации Хинчина [7].

[6] См.: *Опойцев В. И.* Нелинейный закон больших чисел // А и Т. 1994. № 4. С. 65–75.

[7] *Хинчин А. Я.* Симметрические функции на многомерных поверхностях // Сб. памяти А. А. Андронова. М.: Изд. АН СССР, 1955. С. 541–576.

Глава 4

Сходимость

4.1. Разновидности

Вот три основных вида вероятностной сходимости, которые отличаются друг от друга не только по форме, но и по сути.

- *Последовательность случайных величин* X_n *сходится к с. в.* X ***по вероятности***, $X_n \xrightarrow{P} X$, *если для любого* $\varepsilon > 0$

$$\boxed{\mathrm{P}(|X_n - X| > \varepsilon) \to 0 \quad \text{при} \quad n \to \infty.}$$

- *Последовательность случайных величин* X_n *сходится к с. в.* X ***в среднеквадратическом***, $X_n \xrightarrow{\text{с. к.}} X$, *если*

$$\boxed{\mathrm{E}\,(X_n - X)^2 \to 0.}$$

- *Последовательность случайных величин* X_n *сходится к с. в.* X ***почти наверное*** (*синоним*: *«с вероятностью* 1»), $X_n \xrightarrow{\text{п. н.}} X$, *если*

$$\boxed{\mathrm{P}\{|X_k - X| < \varepsilon,\ k \geqslant n\} \to 1 \quad \text{при} \quad n \to \infty.}$$

Вспоминая, что с. в. X_n есть на самом деле функция $X_n(\omega)$, можно сказать так: $X_n \xrightarrow{\text{п. н.}} X$, если $X_n(\omega)$ сходится к $X(\omega)$ в обычном смысле почти для всех ω, за исключением ω-множества нулевой вероятности (меры).

Перечисленные определения обладают общим недостатком — используют предельное значение X, которое не всегда известно. Для преодоления подобной трудности в анализе изобретено понятие

фундаментальной последовательности (*последовательности Коши*). Аналогичный трюк работает и в теории вероятностей.

Последовательность с. в. X_n *называется фундаментальной — по вероятности, в среднем, почти наверное, — если*

$$\mathbf{P}(|X_n - X_m| > \varepsilon) \to 0, \qquad \mathbf{E}(X_n - X_m)^2 \to 0, \qquad \mathbf{P}\{|X_k - X_l| < \varepsilon;\ k, l \geqslant n\} \to 1,$$

при $m, n \to \infty$ *и* $\varepsilon > 0$.

4.1.1 Признак сходимости Коши. *Для вероятностной сходимости* $X_n \to X$ *в любом указанном выше смысле необходима и достаточна фундаментальность последовательности* X_n *в том же смысле.* (?)

Взаимоотношения. Сходимость по вероятности из перечисленных разновидностей самая слабая. Импликация «$\xrightarrow{\text{п. н.}}$» $\Rightarrow$ «$\xrightarrow{P}$» очевидна, а неравенство Чебышева [1] обеспечивает «$\xrightarrow{\text{с. к.}}$» $\Rightarrow$ «$\xrightarrow{P}$».

Обратное в обоих случаях неверно.

◀ Последовательность независимых с. в. X_n при условии

$$\mathbf{P}\{X_n = 0\} = 1 - \frac{1}{n}, \qquad \mathbf{P}\{X_n = n\} = \frac{1}{n}$$

сходится к нулю по вероятности [2], но не сходится ни в среднем, ни почти наверное. Действительно, $\mathbf{E}\,X_n^2 = n \not\to 0$, а расходимость почти наверное следует из леммы Бореля—Кантелли, поскольку

$$\sum^{\infty} \mathbf{P}\{|X_k| > \varepsilon\} = \sum^{\infty} \frac{1}{k} = \infty. \quad \blacktriangleright$$

Стоящие за кадром общие причины достаточно очевидны. Если сходимость по вероятности означает стремление к нулю меры событий $\{|X_n| > \varepsilon\}$, то для «п. н.»-сходимости требуется — достаточно быстрое стремление к нулю этой меры. Понятно, что это разные ситуации.

Для «с. к.»-сходимости само по себе стремление к нулю меры событий $\{|X_n| > \varepsilon\}$ вообще недостаточно, поскольку здесь вступает в игру другой фактор: значения X_n на «плохих траекториях». Поэтому, кстати, «с. к.»-сходимость не следует даже из «п. н.»-сходимости.

---

[1] См. предметный указатель.

[2] Поскольку $\mathbf{P}(|X_n| > \varepsilon) = 1/n \to 0$ при малых ε.

Пример:

$$\mathbf{P}\{X_n = 0\} = 1 - \frac{1}{n^2}, \qquad \mathbf{P}\{X_n = n\} = \frac{1}{n^2}.$$

Для «п. н.»-сходимости стремление к нулю $\mathbf{P}\{|X_n| > \varepsilon\}$ достаточно быстрое[3)], но $\mathbf{E}\, X_n^2 = 1 \not\to 0$.

Наконец, $X_n \xrightarrow{\text{с. к.}} 0$ в случае

$$\mathbf{P}\{X_n = 0\} = 1 - \frac{1}{n}, \qquad \mathbf{P}\{X_n = 1\} = \frac{1}{n},$$

но X_n не сходится к нулю почти наверное.

Итак, $\left\{ \boxed{\xrightarrow{\text{п. н.}}} \atop \boxed{\xrightarrow{\text{с. к.}}} \right\} \Rightarrow «\xrightarrow{P}»$. Других импликаций нет.

Заметим, что для последовательностей случайных векторов модуль заменяется нормой — без каких бы то ни было иных изменений, как в определениях, так и в выводах.

Упражнения

- Если X_n — последовательность *независимых* случайных величин и $X_n \xrightarrow{P} X$, то $\mathbf{P}\{X = \mathbf{E}\, X\} = 1$.
- Для сходимости $X_n \xrightarrow{\text{п. н.}} X$ необходимо и достаточно

$$\lim_{k\to\infty} \mathbf{P}\left\{\sup_{n>k} |X_n - X| > \varepsilon\right\} = 0$$

при любом $\varepsilon > 0$.

- Если $X_n \xrightarrow{P} X$, то существует подпоследовательность $X_{n_k} \xrightarrow{\text{п. н.}} X$.
- Если $X_n \xrightarrow{P} X$ и X_n, X ограничены, то $X_n \xrightarrow{\text{с. к.}} X$.
- Пусть монотонная последовательность неотрицательных случайных величин,

$$0 \leqslant X_1 \leqslant X_2 \leqslant \ldots,$$

имеет равномерно ограниченные матожидания $\mathbf{E}\{X_n\} < m < \infty$. Тогда

$$X_n \xrightarrow{\text{п. н.}} X \quad \text{и} \quad \mathbf{E}\{X_n\} \to \mathbf{E}\{X\} < \infty.$$

[3)] $\sum^{\infty} \mathbf{P}\{|X_k| > \varepsilon\} = \sum^{\infty} \frac{1}{k^2} < \infty$, что для обоснования «п. н.»-сходимости позволяет задействовать лемму Бореля—Кантелли.

4.2. Сходимость по распределению

Есть еще одна важная разновидность вероятностной сходимости, которая слабее предыдущих, и потому — шире применима.

Последовательность случайных величин X_n сходится к с. в. X ***по распределению***, $X_n \xrightarrow{D} X$, *если последовательность соответствующих функций распределения $F_n(x)$ слабо сходится к функции распределения $F(x)$.*

Слабая сходимость $F_n(x) \xrightarrow{w} F(x)$ означает

$$\int_{-\infty}^{\infty} \phi(x)\, dF_n(x) \to \int_{-\infty}^{\infty} \phi(x)\, dF(x),$$

т. е.

$$\mathbf{E}\{\phi(X_n)\} \to \mathbf{E}\{\phi(X)\} \tag{4.1}$$

для любой *непрерывной и ограниченной* функции $\phi(x)$. Это равносильно *поточечной сходимости $F_n(x) \to F(x)$ в точках непрерывности $F(x)$.*

Последнее утверждение является, вообще говоря, теоремой, — справедливой в силу монотонности и ограниченности функций распределения. Если $F(x)$ непрерывна и $F_n(x) \xrightarrow{w} F(x)$, то эта сходимость равномерная, — опять-таки по причине «монотонности и ограниченности». По той же самой причине множество Im всех функций распределения слабо предкомпактно, т. е. из любой последовательности $F_n(x)$ можно выделить слабо сходящуюся подпоследовательность, но не обязательно к функции из Im (за подробностями можно обратиться к [3, 9]).

Импликация «$\xrightarrow{P}$» $\Rightarrow$ «$\xrightarrow{D}$» очевидна. Обратное неверно. Например, пусть речь идет о бросании симметричной монеты, и при четном n с. в. $X_n = 1$, если выпадает герб, и $X_n = 0$ в противном случае, а при нечетном n — $X_n = 0$, если выпадает герб, и $X_n = 1$ в противном случае. Сходимость по распределению есть, по вероятности — нет.

Но $\boxed{\text{если} X_n \xrightarrow{D} 0,\ \text{то}\ X_n \xrightarrow{P} 0}$. (?)

4.2.1 Теорема. *Сходимость по распределению $X_n \xrightarrow{D} X$ равносильна равномерной (на любом конечном промежутке) сходимости $\varphi_n(\lambda) \to \varphi(\lambda)$ характеристических функций.*

◀ Импликация

$$X_n \xrightarrow{D} X \quad \Rightarrow \quad \varphi_n(\lambda) \to \varphi(\lambda)$$

вытекает из (4.1), если положить $\phi(x) = e^{i\lambda x}$.

Обратно, пусть $\varphi_n(\lambda) \to \varphi(\lambda)$, — тогда

$$\mathbf{E}\{\phi(X_n)\} = \int_{-\infty}^{\infty} \widehat{\phi}(\lambda)\varphi_n(\lambda)d\lambda \to \int_{-\infty}^{\infty} \widehat{\phi}(\lambda)\varphi(\lambda)d\lambda = \mathbf{E}\{\phi(X)\},$$

где $\widehat{\phi}(\lambda)$ — преобразование Фурье функции $\phi(\lambda)$. ▶

Конечно, это лишь набросок доказательства. Углубиться в детали можно в любом курсе теории вероятностей [3, 9, 20, 31]. Теорема 4.2.1 часто используется, значительно облегчая суммирование случайных величин.

4.3. Комментарии

Различные виды вероятностной сходимости задают игровое поле для многочисленных постановок задач [19], большинство которых не имеют никакого прикладного значения, но это и не требуется. Задачи способствуют всестороннему изучению предмета. Тренировка, расширение кругозора, познание внутренних механизмов, постановка новых вопросов, — вот, собственно, их главная роль.

Беда в том, как показывает опыт, что разновидности вероятностной сходимости часто остаются «вещью в себе», устроенной по формально понятным правилам, но при отсутствии удобных мысленных образов. Путеводной нити в результате — нет, и задачи приходится решать «наощупь». В такого рода ситуациях, чтобы уменьшить ощущение хаоса, полезно отталкиваться от примеров.

Вот вопрос, который значительную часть населения ставит в тупик. Существует ли последовательность событий $A_1, A_2, \ldots$ такая, что $\mathbf{P}\{A_k\} \to 1$ при $n \to \infty$, но

$$\mathbf{P}\left\{\bigcap_{k=n}^{\infty} A_k\right\} = 0 \quad \text{при любом } n?$$

Положительный ответ дает последовательность одинаково ориентированных дуг A_k длины $k/(k+1)$ на единичной окружности Ω, у которых начало следующей (по номеру) дуги совмещается с концом предыдущей.

Другая проблема. Случайная величина X_n, определяемая соотношениями

$$\mathbf{P}\{X_n = 0\} = 1 - \frac{1}{n}, \qquad \mathbf{P}\{X_n = n^2\} = \frac{1}{n}$$

сходится к нулю по вероятности, $X_n \xrightarrow{P} 0$, но $\mathbf{E}\{X_n\} \to \infty$.

Так часто бывает в некоторых типах игр, в том числе — биржевых. Ожидаемый выигрыш, как говорится, «выше крыши», на деле — почти гарантированный проигрыш.

Сходимость матожиданий. Последняя «неприятность» подчеркивает принципиальную роль утверждений, в которых к вероятностной сходимости можно добавить сходимость матожиданий. Вот несколько полезных и достаточно простых (для обоснования) утверждений.

- *Пусть* $X_n \xrightarrow{\text{п. н.}} X$ *и все* $|X_n| < Y$, *где с. в.* Y *имеет конечное матожидание. Тогда* X *тоже имеет конечное матожидание и* $\mathbf{E}\{X_n\} \to \mathbf{E}\{X\}$.
- *Пусть* $X_n \xrightarrow{\text{с. к.}} X$ *и все* $|X_n| < \infty$. *Тогда* $\mathbf{E}\{|X|\} < \infty$ *и* $\mathbf{E}\{X_n\} \to \mathbf{E}\{X\}$.

Последовательность X_n называется *равномерно интегрируемой*, если

$$\sup_n \int\limits_{|x|>M} |x|\, dF_n(x) \to 0 \quad \text{при} \quad M \to \infty,$$

где $F_n(x)$ — функция распределения X_n.

Условие равномерной интегрируемости часто выполняется, и это ликвидирует массу возможных неприятностей. *В условиях равномерной интегрируемости* X_n:

(i) $\sup_n \mathbf{E}\{|X_n|\} < \infty$. (?)

(ii) *Из* $X_n \xrightarrow{D} X$ *следует существование* $\mathbf{E}\{X\}$ *и* $\mathbf{E}\{X_n\} \to \mathbf{E}\{X\}$. (?)

Без равномерной интегрируемости ситуация менее благоприятна. Пусть, например, X имеет распределение Коши, а

$$X_n = \begin{cases} X, & \text{если } |X| \leqslant n, \\ 0, & \text{если } |X| > n. \end{cases}$$

Тогда $X_n \xrightarrow{D} X$, все матожидания $\mathbf{E}\{X_n\}$ существуют, но $\mathbf{E}\{X\} = \infty$.

Если же $X_n = X/n + Z$, где $\mathbf{E}\{Z\} < \infty$, — получается другая «неприятность»:

$$X_n \xrightarrow{D} Z, \quad \mathbf{E}\{Z\} < \infty,$$

но ни одно $\mathbf{E}\{X_n\}$ не существует.

4.4. Закон «нуля или единицы»

Лемма Бореля—Кантелли служит простейшей иллюстрацией действия механизма, исключающего из рассмотрения все вероятности за исключением крайних. Обзор существенно расширяет *колмогоровский закон* [13] «нуля или единицы», утверждающий следующее.

4.4.1 Теорема. *Если $X_1, X_2, \ldots$ — независимые случайные величины, а событие A определяется поведением только бесконечно далекого хвоста последовательности $X_1, X_2, \ldots$ и не зависит от значений $X_1, \ldots, X_n$ при любом конечном n, — то*

$$\textit{либо } \mathbf{P}\{A\} = 0, \quad \textit{либо } \mathbf{P}\{A\} = 1. \tag{4.2}$$

События, зависящие только от «хвоста», называют *остаточными*. Таковы, например, *события*: *сходимости ряда* $\sum_{k=1}^{\infty} X_k$ *либо самой последовательности* X_k; *ограниченности верхнего предела* $\overline{\lim_{k\to\infty}} X_k < \infty$ и т. п.

◀ Идея доказательства проста[4]. Остаточное событие A — это некоторое множество A бесконечных траекторий $\xi_1, \xi_2, \ldots$, которое может быть аппроксимировано множеством A_ε конечных траекторий $\xi'_1, \ldots, \xi'_m$. «Аппроксимировано» — в смысле $\mathbf{P}\{A \triangle A_\varepsilon\} < \varepsilon$ при задании подходящего $m(\varepsilon)$.

В силу остаточности, A не зависит от A_ε, т. е.

$$\mathbf{P}\{A \cap A_\varepsilon\} = \mathbf{P}\{A\} \cdot \mathbf{P}\{A_\varepsilon\}, \tag{4.3}$$

а поскольку

$$|\mathbf{P}\{A\} - \mathbf{P}\{A_\varepsilon\}| \leqslant \mathbf{P}\{A \triangle A_\varepsilon\} < \varepsilon$$

и

$$|\mathbf{P}\{A \cap A_\varepsilon\} - \mathbf{P}\{A\}| \leqslant \mathbf{P}\{A \triangle A_\varepsilon\} < \varepsilon,$$

то (4.3) при $\varepsilon \to 0$ переходит в

$$\mathbf{P}\{A\} = [\mathbf{P}\{A\}]^2,$$

что возможно лишь в случае (4.2). ▶

[4] Но уточнение деталей, особенно с учетом приготовлений, обычно смазывает картину.

4.5. Случайное блуждание

Пусть $X_1, X_2, \ldots$ — независимые с. в., принимающие два значения 1 и -1,

$$\mathbf{P}\{X_k = 1\} = p, \quad \mathbf{P}\{X_k = -1\} = 1 - p.$$

Поведение суммы $S_n = X_1 + \ldots + X_n$ часто интерпретируют как *случайное блуждание*, имея в виду движение частицы по целочисленным точкам действительной прямой. Принимает X_k в k-й момент времени значение 1 (-1) — частица сдвигается на единицу вправо (влево) [5].

Возврат частицы в начало координат равносилен, очевидно, событию $\{S_n = 0\}$. Понятно, что возвраты возможны только в четные моменты $n = 2k$. Интуитивно ясно, что в случае $p = 1/2$ типичные траектории бесконечно много раз проходят через нуль, а в случае $p \neq 1/2$ уходят в бесконечность.

В точной формулировке:

$$\mathbf{P}\{S_n = 0 \quad \textit{б. ч. р.}\} = \begin{cases} 0, & \textit{если } p \neq \dfrac{1}{2}, \\ 1, & \textit{если } p = \dfrac{1}{2}, \end{cases}$$

где б. ч. р. означает «бесконечное число раз».

◀ Легко видеть (опираясь на формулу Стирлинга), что

$$\mathbf{P}\{S_{2k} = 0 \quad \text{б. ч. р.}\} = C_{2k}^k p^n (1-p)^n \sim \frac{[4p(1-p)]^n}{\sqrt{n}}.$$

Поэтому $\sum_{\infty} \mathbf{P}\{S_{2k} = 0\} < \infty$ и $\mathbf{P}\{S_n = 0 \quad \text{б. ч. р.}\} = 0$ в случае $p \neq 1/2$ следует из леммы Бореля—Кантелли.

Что касается ситуации $p = 1/2$, то здесь $\sum^{\infty} \mathbf{P}\{S_{2k} = 0\} = \infty$, но лемма Бореля—Кантелли не работает, поскольку события $\{S_n = 0\}$ не независимы, а колмогоровский закон «нуля или единицы» не применим, потому что событие $\{S_n = 0 \quad \text{б. ч. р.}\}$ не является остаточным. Но доказательство может быть завершено с помощью дополнительных ухищрений [3, 31].

[5] С тем же успехом можно говорить о бросании монеты или о любой другой реализации схемы Бернулли. Определенную популярность имеют игровые интерпретации (задачи о разорении).

Например, $\{S_{2k} = 0 \quad \text{б. ч. р.}\}$ включает в себя *остаточные* события

$$A = \left\{ \overline{\lim} \frac{S_n}{\sqrt{n}} = \infty, \ \underline{\lim} \frac{S_n}{\sqrt{n}} = -\infty \right\} \qquad \text{и} \qquad A_\nu = \left\{ \overline{\lim} \left| \frac{S_n}{\sqrt{n}} \right| > \nu \right\},$$

к которым применим закон «нуля или единицы». Альтернатива $\mathbf{P}\{A_\nu\} = 0$ исключена в силу теоремы 3.1.1. А поскольку $A_\nu \to A$ при $\nu \to \infty$, то и $\mathbf{P}\{A\} = 1$ и, как следствие, $\mathbf{P}\{S_{2k} = 0 \quad \text{б. ч. р.}\} = 1$, поскольку

$$A \subset \{S_{2k} = 0 \quad \text{б. ч. р.}\}. \quad \blacktriangleright$$

Другая, совсем простая, идея доказательства описана в разделе 4.8. Еще одна принципиальная идея может опираться на закон «нуля или единицы» *Хьюита—Сэвиджа*[6]:

Пусть $X_1, X_2, \ldots$ — независимые, одинаково распределенные случайные величины. Тогда вероятность любого события, инвариантного относительно перестановок конечного числа членов $X_1, X_2, \ldots$, — равна нулю или единице[7].

Событие $\{S_n = 0 \quad \text{б. ч. р.}\}$ удовлетворяет такому условию «нечувствительности к конечным перестановкам», поэтому его вероятность равна или 0, или 1, после чего выбор (когда *нуль*, когда *единица*) осуществляется достаточно легко.

Многомерное блуждание. Пусть речь идет о блуждании частицы по двумерной целочисленной решетке. Движения *влево/вправо* и *вверх/вниз* независимы и происходят (каждое) с вероятностью $1/2$.

Вероятность возвращения в нуль через $2n$ шагов равна, очевидно,

$$\mathbf{P}\{S_{2n} = 0\} = \left[C_{2n}^n \left(\frac{1}{2} \right)^{2n} \right]^2 \sim \frac{1}{\sqrt{n}}.$$

Поэтому

$$\sum^{\infty} \mathbf{P}\{S_{2k} = 0\} = \infty.$$

Далее, с теми же ухищрениями, что и выше, вероятность бесконечного числа возвращений в начало координат получается равной 1, что несколько неожиданно, поскольку обнуление координат теперь должно происходить одновременно.

---

[6] *Hewitt E., Savage L. J.* // Trans. Amer. Math. Soc. 1955. **80**. № 2. P. 470–501.

[7] Доказательство можно найти в [3, 31].

При трехмерном блуждании

$$\mathbf{P}\{S_{2n}=0\}=\left[C_{2n}^{n}\left(\frac{1}{2}\right)^{2n}\right]^{3}\sim n^{-3/2},$$

и тогда

$$\sum^{\infty}\mathbf{P}\{S_{2k}=0\}<\infty,$$

что принципиально меняет картину асимптотического поведения. Вероятность возврата становится дробной, а число возвращений на типичных траекториях конечным.

Качественное отличие поведения случайных траекторий в размерностях 2 *и* 3 *часто служит поводом для удивления и некоторого философствования. Циник бы, конечно, не преминул заметить, что с тем же успехом можно удивляться сходимости ряда* $\sum n^{-2}$ *и расходимости* $\sum n^{-1}$. *Возражать, по сути, было бы трудно, хотя удивление — очень ценная вещь*[8]. *Но проще, и продуктивнее, удивляться существованию этого мира, добиваясь понимания по мелочам.*

В то же время надо признать, что новые содержательные интерпретации тривиальных фактов нередко обнаруживают пропасти.

4.6. Сходимость рядов

В классическом анализе сходимость ряда $\sum^{\infty} a_k$ означает варианты

$$A_n=\sum_{1}^{n}a_k,$$

рáвно как сходимость некоторой последовательности A_n равносильна сходимости ряда $\sum^{\infty}(A_k-A_{k-1})$. Несмотря на эквивалентность языков — каждый имеет свои преимущества и недостатки. При изучении случайных последовательностей и рядов возникает аналогичная картина.

4.6.1 Теорема. *Если* $X_1, X_2, \ldots$ *— независимые случайные величины с нулевыми матожиданиями, то для сходимости ряда* $\sum^{\infty} X_k$ *почти*

[8] Лев Толстой жаловался: «Писáть стало трудно — кончается энергия заблуждения».

наверное достаточно сходимости числового ряда:

$$\sum^{\infty} \mathbf{D}\{X_k\} < \infty. \tag{4.4}$$

А если все X_k *ограничены*, $\mathbf{P}\{|X_k| < M\} = 1$, *то условие* (4.4) *и необходимо.*

◀ Из неравенства Колмогорова следует

$$\mathbf{P}\Big\{\sup_{n>k} |S_n - S_k| > \varepsilon\Big\} = \lim_{n\to\infty} \mathbf{P}\Big\{\max_{k<m<n} |S_m - S_k| > \varepsilon\Big\} \leqslant \frac{1}{\varepsilon^2} \sum_{m>k} D_{X_m} \to 0$$

при $k \to \infty$, что в итоге обеспечивает достаточность. С необходимостью возни несколько больше [21, 31]. ▶

Разумеется, если X_k в теореме 4.6.1 имеют ненулевые матожидания m_{X_k}, то все остается в силе при дополнительном предположении о сходимости ряда $\sum^{\infty} m_{X_k}$. (?)

Специфика случайных рядов (в отличие от последовательностей общего вида) проявляется в следующем полезном факте.

4.6.2 Теорема. *Если* $X_1, X_2, \ldots$ *— независимые случайные величины, то для ряда* $\sum^{\infty} X_k$ *понятия сходимости почти наверное по вероятности и по распределению — эквивалентны* [29].

Если $\mathbf{P}\{|X_k| > \varepsilon_k\} < \delta_k$ и ряды $\sum^{\infty} \varepsilon_k$, $\sum^{\infty} \delta_k$ сходятся, то $\sum^{\infty} X_k$ сходится п. н. (?)

4.7. Предельные распределения

При делении на n сумма

$$S_n = X_1 + \ldots + X_n$$

сходится в том или ином смысле к матожиданию $\mu = \mathbf{E}\{X_k\}$. Специальная «нормировка»

$$\widehat{S}_n = \frac{S_n - n\mu}{\sqrt{n}}$$

позволяет стабилизировать среднеквадратическое отклонение $\widehat{S}_n$, и под этим «микроскопом» детально изучать поведение $\widehat{S}_n$.

4.7.1 Теорема. *Пусть $X_1, X_2, \ldots$ — независимые случайные величины, имеющие одинаковое распределение со средним $\mu = 0$ и дисперсией σ^2. Тогда с. в. $S_n/\sqrt{n}$ сходится по распределению к с. в. S, имеющей нормальное распределение с нулевым матожиданием и дисперсией σ^2.*

◀ Пусть $\varphi(\lambda)$ — характеристическая функция X_k. В силу (2.12)

$$\mathbf{E}\left(\exp\left\{i\lambda\frac{S_n}{\sqrt{n}}\right\}\right) = \left[\varphi\left(\frac{\lambda}{\sqrt{n}}\right)\right]^n.$$

Разложение х. ф. в ряд (2.13) дает

$$\varphi(\lambda) = 1 - \frac{\sigma^2\lambda^2}{2} + o(\lambda^2),$$

откуда при $n \to \infty$

$$\left[\varphi\left(\frac{\lambda}{\sqrt{n}}\right)\right]^n = \left[1 - \frac{\sigma^2\lambda^2}{2n} + o\left(\frac{\lambda^2}{n}\right)\right]^n \to \exp\left\{-\frac{\sigma^2\lambda^2}{2}\right\},$$

т. е. характеристическая функция с. в. $S_n/\sqrt{n}$ сходится к х. ф. нормального закона $\mathcal{N}(0, \sigma^2)$, и по теореме 4.2.1 сама с. в. $S_n/\sqrt{n}$ сходится по распределению к нормальному закону. ▶

Результаты типа теоремы 4.7.1 называют *центральными предельными теоремами*. В приведенном варианте безболезненно можно отказаться от предположения об одинаковости распределения величин X_k. Дальнейшие обобщения связаны с некоторыми нюансами.

Пусть

$$m_k = \mathbf{E}\{X_k\}, \quad \sigma_k^2 = \mathbf{D}\{X_k\}, \quad B_n^2 = \sum_{k=1}^{n} \sigma_k^2.$$

Тогда слабую сходимость

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{S_n - \mathbf{E}\,S_n}{\sqrt{\mathbf{D}\,S_n}} < x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-s^2/2}\,ds = \Phi(x)$$

обеспечивает *условие Ляпунова*: для некоторого $\delta > 0$

$$\frac{1}{B_n^{2+\delta}}\sum_{k=1}^{n} \mathbf{E}\,|X_k - m_k|^{2+\delta} \to 0 \quad \text{при} \quad n \to \infty,$$

а также более свободное *условие Линдеберга*: для любого τ

$$\frac{1}{B_n^2} \int\limits_{|x-m_k| \geqslant \tau B_n} (x-m_k)^2 \, dF_k(x) \to 0 \quad \text{при} \quad n \to \infty,$$

где $F_k(x)$ — функция распределения X_k.

Дополнительные результаты см. в [10].

4.8. Задачи и дополнения

- Если с. в. X распределена по Пуассону с параметром a, то случайная величина $(X-a)/\sqrt{a}$ имеет в пределе (при $a \to \infty$) стандартное нормальное распределение. (?)

- Пусть в задаче о случайном блуждании P_k обозначает вероятность попадания частицы в начало координат из положения $x = k$. Если частица движется вправо с вероятностью p, то

$$P_1 = pP_2 + 1 - p.$$

При этом, как легко сообразить, $P_2 = P_1^2$. Поэтому $P_1 = pP_1^2 + 1 - p$. Решение квадратного уравнения дает два корня: 1 и $(1-p)/p$. В случае $p = 1/2$ оба корня равны 1, т. е. $P_1 = 1$, откуда вытекает $P_k = 1$ при любом k, что означает бесконечное число возвратов частицы в нуль (или любую другую точку).

В общем случае $P_k = 1$, если $p \leqslant 1/2$, и $P_k = (1-p)^k/p^k$, если $p > 1/2$. Это можно интерпретировать как решение *задачи о разорении*, P_k — вероятность проигрыша игроком в сумме k партий при игре против казино с неограниченным ресурсом.

Если же речь идет об игре двух игроков A и B, первый из которых выигрывает отдельные партии с вероятностью $p > 1/2$ и располагает капиталом для проигрыша m партий, а второй — n партий, то A разоряется [9] с вероятностью

$$p_A = \frac{1 - [(1-p)/p]^m}{1 - [(1-p)/p]^{m+n}}. \quad (?)$$

Вероятность разорения второго: $p_B = 1 - p_A$.

При $p \to 1/2$

$$p_A \to \frac{m}{m+n}.$$

- Пусть не равные тождественно нулю с. в. $X_1, X_2, \ldots$ имеют *нулевые математические ожидания, независимы и одинаково распределены*. Тогда суммы $S_n = X_1 + \ldots + X_n$ обладают свойством

$$\mathbf{P}\{\limsup_{n\to\infty} S_n = +\infty\} = \mathbf{P}\{\liminf_{n\to\infty} S_n = -\infty\} = 1, \quad (?)$$

которое при неодинаковой распределенности X_k может нарушаться.

[9] Первым проигрывает m партий.

- **Устойчивые законы.** Изучение сумм независимых с. в. привело к постановке следующего типа вопросов. Если взвешенные суммы слабо сходятся, то что можно сказать о предельном распределении $F(x)$? Одно из направлений возможного ответа — устойчивость $F(x)$.
Распределение $F(x)$ называется *устойчивым*, если независимые с. в. X, Y, распределенные по этому закону, — при сложении, после предварительной подходящей «перенормировки», дают величину, распределенную по тому же закону. Иными словами, найдутся константы, при которых

$$Z = \nu(aX + bY - \mu)$$

имеет распределение $F(x)$.
Точнее и проще говоря, распределение $F(x)$ устойчиво, если при любых a_k и $b_k > 0$ существуют такие a и $b > 0$, что

$$F\left(\frac{x - a_1}{b_1}\right) * F\left(\frac{x - a_2}{b_2}\right) = F\left(\frac{x - a}{b}\right),$$

где звездочка обозначает свертку.
Следующий результат принадлежит *Леви*: *Если* $X_1, X_2, \ldots$ — *независимые одинаково распределенные с. в., и при подходящих* a_k *и* $b_k > 0$ *суммы*

$$\frac{X_1 + \ldots + X_k - a_k}{b_k}$$

слабо сходятся к невырожденному распределению $F(x)$, *то* $F(x)$ — *устойчиво.*

- **Безгранично делимые законы.** Распределение $F(x)$ называют *безгранично делимым*, если при любом целом k существует такая функция распределения $F_k(x)$, что k-кратная свертка $F_k(x)$ дает $F(x)$,

$$F(x) = F_k(x) * \ldots * F_k(x),$$

т. е. корень из характеристической функции $F(x)$ любой k-й степени — оказывается тоже характеристической функцией некоторого закона.
Например, х. ф. $\varphi(\lambda) = e^{-a|\lambda|}$ распределения Коши после извлечения корня дает х. ф. того же распределения Коши с параметром a/k. Поэтому распределение Коши безгранично делимо. То же самое имеет место в отношении нормального, пуассоновского, показательного и ряда других законов.
В общем случае х. ф. безгранично делимого закона обязана быть представимой в каноническом виде Леви—Хинчина:

$$\varphi(\lambda) = \exp\left\{\int_{-\infty}^{\infty} \frac{1 + x^2}{x^2}\left(e^{i\lambda x} - \frac{i\lambda x}{1 + x^2} - 1\right) d\mu(x) + i\lambda\xi\right\},$$

при некотором вещественном ξ и неубывающей ограниченной функции $\mu(x)$.
При соблюдении аккуратности терминологии можно сказать следующее. Безгранично делимые законы представляют собой в точности совокупность возможных предельных распределений при суммировании независимых с. в. Тематика устойчивых и безгранично делимых законов становится интересной, когда на теории вероятностей свет начинает сходиться клином. До этого момента обычно есть масса других точек концентрации внимания.

- Центральная предельная теорема «заслоняет свет», и многие часто думают, что «суммы всегда сходятся к нормальному закону». Разумеется, это не так. Вот простой пример.
  Пусть n единичных масс *равномерно* распределены на $[-n, n]$. На единичную массу в начале координат действует гравитационная сила

$$f_n = \sum \frac{\operatorname{sign}(X_k)}{X_k^2}.$$

  В силу равномерного распределения X_k,

$$\mathbf{E}\left[\exp\left\{i\lambda\frac{\operatorname{sign}(X_k)}{X_k^2}\right\}\right] = \int_{-n}^{n} \exp\left\{i\lambda\frac{\operatorname{sign}(x)}{x^2}\right\}\frac{dx}{2n} = \frac{1}{n}\int_0^n \cos\left(\frac{\lambda}{x^2}\right)dx.$$

  В итоге (детали см. в [15])

$$\mathbf{E}\{e^{i\lambda f_n}\} \to e^{-c|\lambda|^{1/2}} \quad \text{при} \quad n \to \infty.$$

  Соответствующее предельное распределение в элементарных функциях не выражается.

- **Мартингалы.** При изучении сходимости особую роль играют последовательности случайных величин X_n, удовлетворяющие условию

$$\mathbf{E}\{X_{n+1}|X_1, \ldots, X_n\} = X_n$$

  и называемые *мартингалами*.
  В случае

$$\mathbf{E}\{X_{n+1}|X_1, \ldots, X_n\} \leqslant X_n$$

  последовательность с. в. X_n называют *полумартингалом*. Полумартингалы для процедур типа стохастической аппроксимации могут служить как раз аналогами функций Ляпунова (см. предыдущий раздел).
  Теория мартингалов довольно обширна [8, 16]. Ее эффективность определяется простым фактом:

  Теорема. *Мартингал X_n с равномерно ограниченными моментами $\mathbf{E}\{X_n^2\}$ сходится почти наверное.*

Глава 5

Марковские процессы

5.1. Цепи Маркова

В детерминированном случае широко распространены динамические процессы вида [1]

$$x_{n+1} = f(x_n),$$

о которых можно было бы сказать так. Какова бы ни была последовательность $x_1, \ldots, x_n$, будущее развитие процесса (при $t > n$) зависит только от x_n.

Вероятностный аналог этого утверждения служит определением *процесса Маркова*, каковым называют *последовательность случайных величин (векторов)* $X_1, \ldots, X_n, \ldots$, *в которой «будущее»* $X_{t>n}$ *определяется только величиной* X_n *и не зависит от предыстории* $X_1, \ldots, X_{n-1}$.

При этом подразумевается зависимость *распределений* с. в. X_{n+1} от X_n (а также от n — в нестационарном случае), и речь идет о динамике условных плотностей распределения $\rho(X_{n+1}|X_n)$.

Относительно конкретных траекторий $x_1, \ldots, x_n, \ldots$ можно говорить, что x_{k+1} есть реализация случайной величины X_{k+1}, имеющей распределение $\rho(X_{k+1}|X_k = x_k)$.

Простейший пример цепи Маркова:

$$S_1, \ldots, S_n, \ldots,$$

где $S_n = Y_1 + \ldots + Y_n$, а Y_n — n-й член случайной «нуль-один»-последовательности в схеме Бернулли. Очевидно,

$$S_{n+1} = S_n + Y_n,$$

откуда ясно, что S_t при $t > n$ зависит только от $S_{t=n}$ и не зависит от предыстории $S_{t<n}$.

[1] Либо $x_{n+1} = f(x_n, n)$ в неавтономном варианте.

Широкий класс марковских процессов дают процедуры адаптивной подстройки параметров вида

$$c_{k+1} = \varphi_k(c_k, \xi_k), \tag{5.1}$$

где ξ_k — измеряемый шумящий, а c_k — настраиваемый параметр[2].

Пример испытаний Бернулли порождает двойственное чувство. С одной стороны — облегчение, поскольку выясняется, что речь идет о простых вещах. С другой — непонятно, зачем городить огород, когда со случайным блужданием и так можно разобраться.

Это принципиальный момент. Общие схемы всегда связаны с «головной болью». Скажем, механическую задачу часто проще решить, не переводя ее в гамильтонову форму. Стандарты порождают дополнительные проблемы, вынуждая тратить силы на канонизацию. В то же время абстрактные модели приводят разнообразные объекты в соприкосновение, взаимно обогащая их. Сведения о случайном блуждании становятся полезны совсем для других содержательных задач, если те укладываются в прокрустово ложе марковского процесса. И постепенно — развитие общей теории проливает свет на разнообразие частных случаев. Стоит убедиться, что задача «укладывается», как начинает работать весь арсенал уже готовых методов и фактов.

Переходные вероятности. Марковский процесс с дискретным временем и *счетным* пространством состояний называют *марковской цепью*. Как правило, подразумевается следующая модель. Состояния пронумерованы. Система (частица), находясь в k-й момент времени в j-м состоянии в $(k+1)$-й момент попадает в i-е состояние с вероятностью P_{ij}, и тогда при распределении частицы по состояниям с вероятностями p_j^k в следующий момент получается распределение

$$p_i^{k+1} = \sum_j P_{ij} p_j^k, \tag{5.2}$$

или в векторном виде $\mathbf{p}^{k+1} = P\mathbf{p}^k$, где $P = [P_{ij}]$ называют *матрицей переходных вероятностей*.

[2] Например, $\xi_1, \xi_2, \ldots$ в задаче обучения распознаванию образов может быть обучающей последовательностью, а c_k вектором решающего правила (дискриминантной функции).

Модель (5.1) помещается в рамки данной схемы, если пространство переменных c_k разбить на клетки (состояния) и на базе (5.1) вычислить переходные вероятности.

Помимо описанной интерпретации (частица, «пребывающая» в j-м состоянии с вероятностью p_j^k) речь может идти о множестве большого числа частиц. Динамика каждой — определяется той же матрицей переходных вероятностей P, а p_j^k обозначает долю частиц, находящихся в j-м состоянии в k-й момент.

Данную модель (в которой матрица P не зависит от k) называют еще *однородной цепью Маркова.*

Понятно, что $\mathbf{p}^{k+m} = P^m \mathbf{p}^k$, т. е. динамика распределений $\mathbf{p}^k$ определяется итерациями матрицы P. При этом, очевидно, $P^{n+m} = P^n P^m$, что называют *уравнением Колмогорова—Чепмена.*

В частности, стационарные распределения $\mathbf{p}^*$ оказываются собственными векторами матрицы P, $\mathbf{p}^* = P\mathbf{p}^*$, а сходимость $\mathbf{p}^k \to \mathbf{p}^*$ — одним из центральных вопросов.

В теории марковских процессов большое внимание уделяется классификации состояний. Состояние x_i называют *достижимым* из x_j, если $P_{ij}^k > 0$ при некотором $k > 0$, т. е. существует ненулевая вероятность через некоторое число шагов из j-го состояния попасть в i-е. Состояния, достижимые друг из друга, называют *сообщающимися.*

Если x достижимо из y, но не наоборот, то состояние y называют *несущественным.* Множество всех существенных состояний разбивается на непересекающиеся классы сообщающихся состояний. Если такой класс всего один, — система называется *неразложимой.*

Состояние x_i называют *возвратным*, если вероятность возвращения в x_i равна 1.

Наконец, состояние x_i считается *периодическим*, если наибольший общий делитель (*период состояния*) чисел k, для которых $P_{ii}^k > 0$, — равен $d > 1$.

5.2. Стохастические матрицы

Положительная матрица $P \geqslant 0$ с единичными столбцовыми суммами, $\sum_i P_{ij} = 1$, называется *стохастической.*

Легко видеть, чтобы итерационная процедура (5.2) на каждом следующем шаге порождала нормированное распределение, $\sum_i p_i^{k+1} = 1$, необходимо как раз $\sum_i P_{ij} = 1$.

Если иметь в виду системы с *конечным числом состояний*, то стохастические матрицы изучаются в линейной алгебре, и там, кстати, многие результаты в подходящем контексте воспринимаются достаточно легко и просто. Затевать сыр-бор в рамках **ТВ** вряд ли имеет смысл, проще отослать к [5, т. 3]. Здесь ограничимся перечислением стержневых результатов с привязкой к вероятностной интерпретации.

- Собственный вектор $\mathbf{p}^* \geqslant 0$, отвечающий собственному значению $\lambda = 1$, у стохастической матрицы существует всегда, т. е. всегда существует стационарное распределение $\mathbf{p}^* = P\mathbf{p}^*$.

- Если матрица P строго положительна (все $P_{ij} > 0$) или же $P^k > 0$ при некотором k, то все стационарные вероятности $p_j^* > 0$, причем итерации $\mathbf{p}^k$ сходятся к $\mathbf{p}^* > 0$, а итерации $P^k \to P_\infty$, где у P_∞ все столбцы одинаковы и равны $\mathbf{p}^*$. Процесс в этом случае называют *эргодическим*.

- Условие «$P^k > 0$ при некотором k» необходимо и достаточно для *примитивности* стохастической матрицы, т. е. для того, чтобы спектр P, за исключением ведущего собственного значения $\lambda = 1$, лежал строго внутри единичного круга. Примитивность P означает отсутствие периодических состояний. В случае *импримитивной*[3] матрицы P предел $\mathbf{p}^k$ может не существовать. Но предел имеют средневзвешенные суммы,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} P^k = P_\infty.$$

Неразложимость. Матрица P называется *разложимой* (*неразложимой*), если одинаковой перестановкой строк и столбцов она приводится (не приводится) к виду

$$\begin{bmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{bmatrix},$$

где P_{11} и P_{22} квадратные матрицы.

Иными словами, P неразложима, если не существует такого подмножества индексов J, что $P_{ik} = 0$ для всех $i \in J$, $k \notin J$.

Система уравнений $Px = x$ с разложимой матрицей, по сути, имеет вид

$$\begin{bmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix},$$

т. е.

$$P_{11}\mathbf{x}_1 + P_{12}\mathbf{x}_2 = \mathbf{x}_1,$$

$$P_{22}\mathbf{x}_2 = \mathbf{x}_2.$$

Наличие автономной подсистемы $P_{22}\mathbf{x}_2 = \mathbf{x}_2$, которую можно решать независимо, — характеристическое свойство разложимой матрицы.

[3] Не примитивной.

Неразложимость P равносильна либо неравенству $(I + P)^{n-1} > 0$, либо существованию для любой пары индексов i, j такого k, что $P_{ij}^{(k)} > 0$, где $P_{ij}^{(k)}$ обозначает (ij)-й элемент матрицы P^k. Но отсюда не вытекает существование k, при котором $P^k > 0$. Если же главная диагональ неразложимой матрицы P строго положительна, то $P^{n-1} > 0$.

5.2.1. *Если матрица P неразложима, то $\lambda(P) = 1$ является ведущим собственным значением P алгебраической кратности 1, которому отвечает строго положительный собственный вектор. Других положительных собственных значений и векторов у P нет*[4].

5.3. Процессы с непрерывным временем

Марковские процессы с дискретным временем имеют свой круг приложений, но гораздо более типичны системы, в которых переход из состояния в состояние происходит в случайные моменты времени (поступления заявки, поломки прибора, окончания ремонта).

Ситуация во многом аналогична предыдущей. Система, находясь в нулевой момент времени в j-м состоянии в момент Δt попадает в i-е состояние с вероятностью $P_{ij}(\Delta t)$, и тогда при начальном распределении системы по состояниям с вероятностями $p_j(0)$ в следующий момент получается распределение

$$p_i(\Delta t) = \sum_j P_{ij}(\Delta t) p_j(0), \tag{5.3}$$

или в векторном виде $\mathbf{p}(\Delta t) = P(\Delta t)\mathbf{p}(0)$, а уравнение Колмогорова—Чепмена переходит в[5]

$$P_{ij}(t + s) = \sum_k P_{ik}(t) P_{kj}(s),$$

что является элементарным следствием формулы полной вероятности.

Выбор фиксированного шага $t = n\Delta$ сразу возвращает ситуацию в прежнее русло (с дискретным временем).

---

[4] Но P может иметь другие собственные значения на единичной окружности — со всеми вытекающими отсюда «неприятностями».

[5] Речь идет о стационарном случае, в котором вероятности не меняются при изменении точки отсчета.

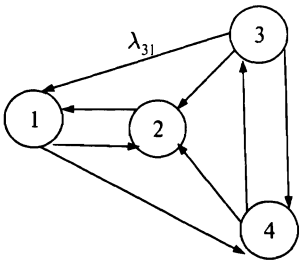


Рис. 5.1

Переходы системы из состояния в состояние удобно мыслить происходящими под воздействием потоков событий (отказов, заявок, восстановлений, запросов, регистраций). Пусть, например, λ_{ij} обозначает интенсивность пуассоновского потока, под воздействием которого система переходит из j-го состояния в i-е с вероятностью $\lambda_{ij}\Delta t + o(\Delta t)$ за время Δt. При этом часто модель сопровождается графом состояний (рис. 5.1), на котором ориентированные дуги между узлами событий отвечают возможным переходам.

Вероятность $p_i(t+\Delta t)$ складывается из двух частей: $\sum\limits_{k\neq i} p_k(t)\lambda_{ik}\Delta t$ — вероятности того, что за время Δt система придет в i-е состояние из других состояний, и вероятности $p_i(t)\Big\{1-\sum\limits_{k\neq i}\lambda_{ki}\Delta t\Big\}$ — того, что система не уйдет из i-го состояния, т. е.

$$p_i(t+\Delta t) = \sum_{k\neq i} p_k(t)\lambda_{ik}\Delta t + p_i(t)\Big\{1 - \sum_{k\neq i}\lambda_{ki}\Delta t\Big\} + o(\Delta t). \qquad (5.4)$$

Если положить

$$\lambda_{ii} = -\sum_{k\neq i}\lambda_{ki},$$

то (5.4) — после переноса $p_i(t)$ влево, деления на Δt и предельного перехода $\Delta t \to 0$ — приводит к *уравнениям Колмогорова*

$$\dot{p}_i(t) = \sum_k \lambda_{ik}p_k(t), \qquad i = 1, \ldots, n.$$

При более внимательном подходе к предмету здесь возникают детали, на которых в случае беглой экскурсии лучше не останавливаться, иначе создается впечатление о наличии сложностей, которых на самом деле нет.

Уравнения Колмогорова в первую очередь используются для определения стационарных решений, для чего приравниваются нулю правые части. Посмотрим, как это делается на примере популярной модели процесса рождения и гибели. Соответствующий граф

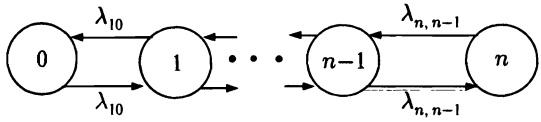


Рис. 5.2

состояний вытянут в цепочку (рис. 5.2). Потоки рождений, переводящие систему из i-го состояния в $(i+1)$-е, имеют интенсивности $\lambda_{i,i+1}$, а процессы гибели, связанные с переходами $i+1 \Rightarrow i$, — интенсивности $\lambda_{i+1,i}$. В моделях массового обслуживания рождению сопоставляется обычно приход заявки (клиента) в систему, гибели — уход обслуженного клиента из системы. В случае радиоактивного распада речь может идти о нейтронах. Генетические модели терминологических пояснений даже не требуют.

Несложные выкладки показывают, что стационарное решение определяется вероятностями:

$$p_k = \frac{\lambda_{k-1,k} \dots \lambda_{12}\lambda_{01}}{\lambda_{k,k-1} \dots \lambda_{21}\lambda_{10}} p_0, \tag{5.5}$$

$$k = 1, \dots, n, \qquad p_0 = \left(1 + \frac{\lambda_{01}}{\lambda_{10}} + \dots + \frac{\lambda_{n-1,n} \dots \lambda_{12}\lambda_{01}}{\lambda_{n,n-1} \dots \lambda_{21}\lambda_{10}}\right)^{-1}.$$

Стандартный вариант в теории массового обслуживания: клиенты в систему поступают с интенсивностью λ, т. е. все $\lambda_{k,k+1} = \lambda$, но прием заявок прекращается при переполнении системы, $\lambda_{k,k+1} = 0$ при $k \geqslant \alpha$. Обслуженные клиенты покидают систему с интенсивностью μ, т. е. все $\lambda_{k,k-1} = \mu$.

Тогда (5.5) при условии нормировки $\sum_0^{\alpha} p_k = 1$ дает

$$p_k = \frac{(1-\rho)\rho^k}{1-\rho^{\alpha+1}}, \qquad \rho = \frac{\lambda}{\mu}, \qquad k < \alpha.$$

5.4. О приложениях

Иллюстрации теории марковских цепей простейшими примерами типа бросания монеты укрепляют мнение большинства о бесполезности предлагаемых моделей. Сложными примерами, с другой

стороны, мало кто интересуется. В результате марковские процессы попадают в нишу обширных, но скучных теорий. С этим, однако, ничего не надо делать, потому что такова реальность. Закономерная скука возникает из-за приведения всех задач к одной схеме. Возможностей для заблуждений почти не остается, а рутина деталей, когда «в принципе» все ясно, — не воодушевляет.

В такой ситуации необходимо лишь признать факты и честно расставить акценты. Область обширна и глубоко проработана, но в общем курсе теории вероятностей для нее достаточно совсем немного места. Чтобы ясно было, о чем речь, и где искать, если потребуется.

Парадоксальный момент при этом заключается в том, что проблематика **ТВ** более чем наполовину укладывается в теорию марковских процессов. Разумеется, способ изложения случайного блуждания — дело вкуса и доброй воли, но ряд областей типа массового обслуживания без идеологии марковости много теряют.

Плюс к этому, есть масса совсем простых задач — абсолютно гробовых до тех пор, пока не приходит мысль использовать схему $\mathbf{p}^{k+1} = P\mathbf{p}^k$. Представим, например, что игра в «орлянку» происходит на четырех монетах, каждая из которых выпадает гербом со своей вероятностью p_k, а какая монета бросается следующей — определяется какой-нибудь схемой, типа изображенной на рис. 5.1, в зависимости от выигрыша или проигрыша в текущей партии. Решение сопутствующих вероятностных вопросов здесь практически невозможно без опоры на $\mathbf{p}^{k+1} = P\mathbf{p}^k$.

Конечно, такая задача представляется надуманной. Но, скажем, в генетике есть масса проблем, которые почти без усилий ложатся в готовые марковские схемы. Например, динамика популяций по группам крови. Здесь, в принципе, все настолько прозрачно, что грубые модели даже не требуют особых пояснений, ложась в рамки $\mathbf{p}^{k+1} = P\mathbf{p}^k$. Беда заключается в другом. Для серьезных продвижений не хватает «смычки». Математик не идет дальше иллюстраций, не будучи готов посвятить часть своей жизни копанию в биологических тонкостях. А биолог ограничивается карикатурными моделями, потому что не хватает математической квалификации.

Глава 6

Случайные функции

6.1. Определения и характеристики

Случайной функцией (с. ф.) называется функция $X(t)$, которая при любом значении аргумента t является случайной величиной. Далее предполагается, что t — время, но, в принципе, t может быть и многомерным параметром. Случайную функцию времени называют также *случайным процессом*, хотя более естественно называть случайными процессами механизмы порождения с. ф. — такие как *марковские процессы*, *стохастические дифференциальные уравнения* и т. п.

Видимая простота определения с. ф обманчива. Неприятности, как правило, выявляются, когда случайные величины $X(t)$ и $X(t+\Delta t)$ приводятся в соприкосновение, т. е. Δt устремляется к нулю. В случае непрерывности с. ф возникают трудности с независимостью $X(t)$ и $X(s)$ при близких $t \neq s$, что порождает определенный дискомфорт.

Избежать неприятностей помогает видоизменение точки зрения. *Случайной функцией* называют функцию двух переменных $X(t,\omega)$, где ω — точка вероятностного пространства Ω, на котором задана та или иная вероятностная мера. Зависимость от случая реализуется при этом каждый раз наступлением исхода $\omega_0 \in \Omega$, при котором фактическое течение процесса описывается *траекторией* $X(t,\omega_0)$, которую называют также *реализацией процесса* или *выборочной функцией*.

Функцию

$$X(t,\omega) = a\cos(2\pi\nu t + \varphi),$$

где $\omega = \{a, \nu, \varphi\}$, можно рассматривать как с. ф. Такая модель, конечно, узка — все реализации (траектории) гармонические. Но

$$X(t,\omega) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n\cos\frac{n\pi t}{l} + b_n\sin\frac{n\pi t}{l}\right),$$

где точку вероятностного пространства $\omega \in \Omega$ определяют последовательности $\{a_k, b_k\}$, включает в рассмотрение все интегрируемые на $[-l, l]$ функции, — и остается задать вероятностную меру на Ω.

Плотность $\rho(x, t)$ случайной функции [1] $X(t)$ определяет распределение значений $X(t)$ в момент t. Разумеется, более полной характеристикой процесса является двумерная плотность $\rho(x_1, x_2, t_1, t_2)$, определяющая распределение значений

$$\{X(t_1), X(t_2)\}$$

в разные моменты времени. Понятно, что еще более полную характеристику дают m-мерные плотности.

Для с. ф естественным образом определяются: *матожидание*

$$m_x(t) = \mathbf{E}\{X(t)\} = \int_{-\infty}^{\infty} x\rho(x, t)\, dx$$

и *корреляционная функция* [2]

$$R_{xx}(t, s) = \mathbf{E}\left\{[X(t) - m_x(t)][X(s) - m_x(s)]\right\},$$

которая при $t = s$ превращается в *дисперсию*

$$D_x(t) = R_{xx}(t, t) = \mathbf{E}\left\{[X(t) - m_x(t)]^2\right\}.$$

Упражнения

- Корреляционная функция с. ф.

$$X(t) = \sum_{n=1}^{N}\left(a_n \cos\frac{n\pi t}{l} + b_n \sin\frac{n\pi t}{l}\right),$$

где случайные величины a_n, b_n не коррелированы,

$$\mathbf{E}\{a_n\} = \mathbf{E}\{b_n\} = 0, \qquad \mathbf{D}\{a_n\} = \mathbf{D}\{b_n\} = \sigma_n^2,$$

определяется равенством

$$R_{xx}(s, t) = \sum_{n=1}^{N} \sigma_n^2 \cos\frac{n\pi(t - s)}{l}. \quad (?)$$

- Пусть $X(t)$ имеет нулевое среднее, принимает значения ± 1, число перемен знака подчиняется закону Пуассона с постоянной λ. Тогда

$$R_{xx}(s, t) = e^{-2\lambda(t-s)}. \quad (?)$$

---

[1] Зависимость от ω подразумевается, но не упоминается, чтобы не загромождать игровое поле.

[2] См. раздел 1.8 по поводу терминологии «ковариация ⇔ корреляция».

6.2. Эргодичность

Стационарные функции. Случайный процесс $X(t)$ стационарен, если его характеристики не меняются при сдвиге по оси времени.

Уточнять сказанное можно различным образом. Требуя, например, независимость от сдвига по оси времени n-мерной плотности распределения

$$\rho(x_1, \ldots, x_n, t_1, \ldots, t_n).$$

В этом случае с. ф. $X(t)$ называют *стационарной в узком смысле.*

Менее жесткий вариант: независимость от сдвига по оси времени условного матожидания и корреляционной функции. В этом случае с. ф. $X(t)$ называют *стационарной в широком смысле.*

В том и другом случае, как легко видеть, матожидание

$$\mathbf{E}\{X(t)\} = m_x = \text{const},$$

а корреляция $R_{xx}(t, s)$ зависит только от разности $t - s$, т. е.

$$R_{xx}(t, s) = R_{xx}(t - s) = R_{xx}(\tau).$$

Соответственно,

$$D_x(t) = R_{xx}(t - t) \equiv R_{xx}(0).$$

Какого рода стационарность подразумевается, обычно ясно из контекста, — и это позволяет обходиться без оговорок.

Эргодичность с. ф. Случайная функция, будучи функцией двух переменных, представляет собой ансамбль траекторий $X(t, \omega)$, индексируемых (с определенными весами) параметром $\omega \in \Omega$. Под эргодичностью X обычно понимают равенство среднего значения X по ансамблю и — среднего по времени. Для стационарного процесса это означает

$$\lim_{T\to\infty} \mathbf{E}\left\{\left[\frac{1}{T}\int_{t_0}^{t_0+T} X(t)\,dt - m_x\right]^2\right\} = 0, \tag{6.1}$$

где t_0 — произвольный момент времени, а $m_x = \mathbf{E}\{X(t)\}$.

Таким образом, в варианте (6.1) речь идет о среднеквадратической сходимости

$$\frac{1}{T}\int_{t_0}^{t_0+T} X(t)\,dt \xrightarrow{\text{с. к.}} m_x \quad \text{при} \quad T \to \infty.$$

Понятно, что такого сорта сходимость представляет собой непрерывный аналог закона больших чисел.

Об эргодичности можно говорить по отношению к любой функции $Y(t) = \varphi[X(t)]$ либо $Y(t_1, \ldots, t_n) = \varphi[X(t_1), \ldots, X(t_n)]$. В частности, — по отношению к корреляционной функции, отталкиваясь от

$$Y(t, s) = [X(t) - m_x][X(s) - m_x].$$

Эргодическое свойство позволяет экспериментально определять матожидание любой стационарной функции $Y(t) = \varphi[X(t)]$ не по множеству реализаций, а по данным одной реализации на достаточно большом промежутке времени T:

$$m_y \approx \frac{1}{T} \int_{t_0}^{t_0+T} Y(t)\, dt.$$

Разумеется, эргодичность «даром не дается». Требуются те или иные предположения. В простейших постановках задачи результат достигается довольно просто. Например,

$$\mathbf{E}\left\{\left[\frac{1}{T}\int_{t_0}^{t_0+T} X(t)\, dt - m_x\right]^2\right\} = \mathbf{E}\left\{\frac{1}{T^2}\int_{t_0}^{t_0+T}\int_{t_0}^{t_0+T} [X(t) - m_x][X(s) - m_x]\, dt\, ds\right\} =$$

$$= \frac{1}{T^2}\int_{t_0}^{t_0+T}\int_{t_0}^{t_0+T} R_{xx}(t - s)\, dt\, ds.$$

Поэтому эргодичность стационарной функции по отношению к матожиданию обеспечивает условие

$$\lim_{T\to\infty} \frac{1}{T^2}\int_{t_0}^{t_0+T}\int_{t_0}^{t_0+T} R_{xx}(t - s)\, dt\, ds = 0, \tag{6.2}$$

которое несложными преобразованиями (см. ниже) сводится к

$$\lim_{T\to\infty} \frac{1}{T}\int_{0}^{T}\left(1 - \frac{\tau}{T}\right) R_{xx}(\tau)\, d\tau = 0. \tag{6.3}$$

◀ Переход от (6.2) к (6.3) осуществляется заменой $\tau = t - s$,

$$I = \int_{t_0}^{t_0+T} \int_{t_0}^{t_0+T} R_{xx}(t-s)\,dt\,ds = \int_{t_0}^{t_0+T} \left[\int_{t_0-s}^{t_0+T-s} R_{xx}(\tau)\,d\tau \right] ds.$$

Обозначая далее

$$\eta(\xi) = \int_0^{\xi} R_{xx}(\tau)\,d\tau$$

и интегрируя по частям, получаем

$$I = \{\eta(t_0+T-s) - \eta(t_0-s)\}s\Big|_{t_0}^{t_0+T} + \int_{t_0}^{t_0+T} \{R_{xx}(t_0+T-s) - R_{xx}(t_0-s)\}s\,ds.$$

Окончательно,

$$I = 2\int_0^T (T-\tau)R_{xx}(\tau)\,d\tau. \quad ▶$$

Более общие задачи в связи с эргодичностью [2] возникают, когда речь заходит о происхождении случайного процесса $X(t)$, который может порождаться, например, стохастическим дифференциальным уравнением или иным механизмом. В этих случаях характеристики $X(t)$ приходится «вытаскивать» из других моделей.

6.3. Спектральная плотность

Преобразование Фурье корреляционной функции стационарного процесса [3)],

$$\widehat{R}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R(\tau)e^{-i\omega\tau}\,d\tau \quad \Leftrightarrow \quad R(\tau) = \int_{-\infty}^{\infty} \widehat{R}(\omega)e^{i\omega\tau}\,d\omega,$$

называют *спектральной плотностью* сигнала $X(t)$.

Обратим внимание, что традиционные обозначения здесь вступают в противоречие. Через ω принято обозначать как точку вероятностного пространства, так и круговую частоту. Но для авральных мер все-таки нет оснований. Из контекста ясно, что имеется в виду.

[3)] Нижний индекс xx далее опущен.

Предположим, что $X(t)$ является стационарной с. ф., эргодичной по отношению к своей корреляционной функции. Тогда

$$\boxed{\widehat{R}(\omega) = \lim_{T\to\infty} \mathbf{E}\{\widehat{R}^*(\omega)\},} \tag{6.4}$$

где

$$\widehat{R}^*(\omega) = \frac{1}{2\pi}\int\limits_{-T/2}^{T/2} R^*(\tau)e^{-i\omega\tau}\,d\tau, \qquad R^*(\tau) = \frac{1}{T}\int\limits_{-T/2}^{T/2}[X(s)-m_x][X(s+\tau)-m_x]\,ds.$$

Предельное соотношение (6.4) справедливо в силу предполагаемой эргодичности:

$$\lim_{T\to\infty} \mathbf{E}\{[R^*(\tau)-R(\tau)]^2\} = 0.$$

Перезапись $\widehat{R}^*(\omega)$ в виде

$$\widehat{R}^*(\omega) = \frac{1}{2\pi T}\int\limits_{-T/2}^{T/2}\left\{\int\limits_{-T/2}^{T/2}[X(s)-m_x][X(t)-m_x]e^{i\omega s}\,ds\right\}e^{-i\omega t}\,dt =$$

$$= \frac{1}{2\pi T}\int\limits_{-T/2}^{T/2}[X(s)-m_x]e^{i\omega s}\,ds\int\limits_{-T/2}^{T/2}[X(t)-m_x]e^{-i\omega t}\,dt$$

указывает на справедливость следующего принципиального соотношения:

$$\boxed{\widehat{R}(\omega) = \lim_{T\to\infty}\frac{2\pi}{T}\mathbf{E}\{|\widehat{A}_T(\omega)|^2\},} \tag{6.5}$$

где $\widehat{A}_T(\omega)$ — преобразование Фурье сигнала $A_T(t) = X_T(t) - m_x$, совпадающего с $X(t)-m_x$ на промежутке $t\in[-T/2, T/2]$ и равного нулю вне этого промежутка.

Важная роль соотношения (6.5) заключается в фиксации взаимосвязи спектра корреляционной функции со спектром самого сигнала $X(t)$.

Простейшие свойства спектральной плотности. Из четности $R_{xx}(\tau)$ вытекает

$$\int\limits_{-\infty}^{\infty} R_{xx}(\tau)\sin\omega\tau\,d\tau = 0.$$

Поэтому

$$\widehat{R}_{xx}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R_{xx}(\tau) \cos \omega\tau \, d\tau.$$

Вещественность и положительность $\widehat{R}_{xx}(\omega)$ вытекают из (6.5).

Широкое распространение в теории распространения волн находит очевидное в данном контексте энергетическое соотношение:

$$D_{xx} = \sigma_x^2 = R_{xx}(0) = \int_{-\infty}^{\infty} \widehat{R}_{xx}(\omega) \, d\omega,$$

увязывающее среднюю мощность случайного сигнала с его спектральной плотностью.

Пример. Корреляционная функция $R_{xx}(\tau) = \sigma_x^2 e^{-\gamma|\tau|}$ имеет спектральную плотность

$$\widehat{R}_{xx}(\omega) = \frac{\sigma_x^2}{\pi} \frac{\gamma^2}{\gamma^2 + \omega^2}. \quad (?)$$

6.4. Белый шум

Стационарный случайный сигнал $X(t)$ с постоянной спектральной плотностью

$$\widehat{R}_{xx}(\omega) \equiv G$$

во всем диапазоне частот от нуля до бесконечности — называют *белым шумом*.

Обратное преобразование Фурье приводит в этом случае к дельтаобразной корреляционной функции

$$R_{xx}(\tau) = G \int_{-\infty}^{\infty} e^{i\omega\tau} \, d\omega = 2\pi G \delta(\tau).$$

Таким образом, корреляционная функция белого шума $R_{xx}(\tau) = 0$ при любом $\tau \neq 0$, т. е. значения сигнала в различные моменты времени $X(t)$ и $X(t+\tau)$ —

всегда некоррелированы. Разумеется, это идеализация. О внутренней противоречивости понятия белого шума свидетельствует также бесконечность дисперсии:

$$R_{xx}(0) = 2G \int_0^\infty d\omega = \infty.$$

Но противоречия здесь, вообще говоря, не страшнее несоизмеримости диагонали квадрата со стороной. Необходимо, конечно, принятие мер, связанных с преодолением достаточно серьезных препятствий, — однако бросать все и заниматься сообща проблемой обоснования вовсе необязательно.

Это извечная проблема. Не только в математике, но и в жизни. Как идти своим путем, чтобы не отвлекаться, и насколько все же поглядывать по сторонам, чтобы не потерять гибкость и поддерживать гармонию?

В математике, правда, черно-белые оттенки этой дилеммы гораздо острее и проще. В любой точке пути — развилка. Обосновывать или идти дальше? Соответственно, две группы исследователей со своими симпатиями и антипатиями. Они обычно друг над другом подтрунивают, не желая согласиться, что нужно и то и другое, — хотя у каждого есть резон бежать за своим зайцем.

6.5. Броуновское движение

Случайная функция $X(t)$ называется *процессом с независимыми приращениями*, если для любых $t_0 < t_1 < \ldots < t_n$ случайные величины $X(t_1) - X(t_0), \ldots, X(t_n) - X(t_{n-1})$ независимы.

Процесс считается *однородным*, если распределение

$$X(t) - X(s)$$

определяется только разностью $t - s$.

Однородный процесс $X(t)$ с независимыми приращениями называют *броуновским движением*, или *винеровским процессом*, если все $X(t_k) - X(t_{k-1})$ распределены нормально[4] со средним 0 и дисперсией $|t_k - t_{k-1}|$.

Описание Эйнштейном броуновского движения опиралось на естественные физические соображения. Если $X(t)$ — координата броуновской частицы в момент времени t, то смещение $X(t) - X(0)$ (для определенности $X(0) = 0$) представляет собой сумму большого числа «мелких» независимых слагаемых

$$X(t) = \sum_k [X(t_k) - X(t_{k-1})],$$

[4] Предполагается также $X(0, \omega) = 0$ почти для всех ω.

и центральная предельная теорема дает основания рассчитывать на нормальное распределение $X(t)$.

Плотность распределения частиц $\rho(x, t)$ при этом подчиняется уравнению

$$\rho(x, t+\tau) = \int_{-\infty}^{\infty} \rho(x-y, t) v(\tau, y)\, dy, \tag{6.6}$$

где $v(\tau, y)$ обозначает долю частиц, переместившихся из x в $x+y$ за время τ.

Разложение (6.6),

$$\rho(x, t) + \tau\rho_t(x, t) + o(\tau) = \int_{-\infty}^{\infty} \left\{ \rho(x, t) - y\rho_x(x, t) + \frac{1}{2} y^2 \rho_{xx}(x, t) + \ldots \right\} v(\tau, y)\, dy,$$

в предположении симметрии $v(\tau, y)$ по y и пропорциональности дисперсии времени τ

$$\int_{-\infty}^{\infty} y^2 v(\tau, y)\, dy = 2D\tau$$

приводит к уравнению диффузии:

$$\boxed{\frac{\partial \rho}{\partial t} = D \frac{\partial^2 \rho}{\partial x^2},} \tag{6.7}$$

решением которого при условии $\rho(x, 0) = \delta(x-y)$ является

$$\rho(x, t) = \frac{1}{4\pi D t} e^{-(x-y)^2/(4Dt)}.$$

Винер придал физическим соображениям строгую форму, основанную на представлении функций $X(t, \omega)$ с помощью счетного множества коэффициентов Фурье [5]:

$$X(t, \omega) = \sum_{n=0}^{\infty} a_n(\omega) \sin\left[\left(n+\frac{1}{2}\right)\pi t\right], \qquad a_n(\omega) = \int_0^1 X(t, \omega) \sin\left[\left(n+\frac{1}{2}\right)\pi t\right] dt,$$

где подразумевается процесс на промежутке $[0, 1]$.

Винеровский процесс занимает в теории случайных функций центральное место по целому ряду причин. В первую очередь потому, что в предположениях

$$X(0) = 0, \qquad \mathbf{E}\{X(t)\} = 0, \qquad \mathbf{D}\{X(t) - X(s)\} = t - s \qquad (s \leqslant t)$$

[5] Если коэффициенты Фурье распределены нормально, то их линейная комбинация тоже распределена нормально.

это единственный непрерывный с вероятностью 1 процесс с независимыми приращениями.

Кроме того, винеровский процесс есть марковский процесс [6] — с переходной плотностью, удовлетворяющей уравнению диффузии (6.7).

6.6. Дифференцирование и интегрирование

Из-за недифференцируемости винеровского процесса обычный аппарат математического анализа в теории случайных функций служит лишь ориентиром. Соответствующие инструменты для изучения *стохастических дифференциальных уравнений* строятся на базе понятия *стохастического интеграла* [18]. На понятийном уровне, однако, изложение вполне можно вести с помощью классических понятий интеграла и производной.

Дифференцирование случайной функции $X(t)$,

$$Y(t) = X'(t),$$

оказывается перестановочно с операцией математического ожидания:

$$\mathbf{E}\left\{\frac{dX(t)}{dt}\right\} = \frac{d\mathbf{E}\{X(t)\}}{dt},$$

что сразу следует из перехода к пределу при $\Delta \to 0$ в очевидном равенстве

$$\mathbf{E}\left\{\frac{X(t+\Delta) - X(t)}{\Delta}\right\} = \frac{\mathbf{E}\{X(t+\Delta)\} - \mathbf{E}\{X(t)\}}{\Delta}.$$

Формула для вычисления корреляционной функции производной $Y(t) = X'(t)$,

$$\boxed{R_{yy}(t, s) = \frac{\partial^2 R_{xx}(t, s)}{\partial t \partial s},} \tag{6.8}$$

получается предельным переходом почти так же просто.

[6] Еще винеровский процесс эквивалентно определяется как гауссовский процесс с нулевым матожиданием и корреляционной функцией $R(t, s) = \sigma^2 \min\{t, s\}$.

В случае стационарного процесса из (6.8) сразу следует

$$\boxed{R_{yy}(\tau) = -\frac{\partial^2 R_{xx}(\tau)}{\partial \tau^2}.}$$

Легко видеть, что спектральная плотность производной сигнала $Y(t) = X'(t)$ равна

$$\boxed{\widehat{R}_{yy}(\omega) = \omega^2 \widehat{R}_{xx}(\omega).}$$

В случае интегрирования

$$Y(t) = \int_0^T g(s,t)X(s)\,ds,$$

где функцию $g(s,t)$ называют *ядром интегрального оператора*[7], характеристики $Y(t)$ определяются по формулам:

$$\mathbf{E}\{Y(t)\} = \int_0^T g(s,t)\mathbf{E}\{X(s)\}\,ds,$$

$$R_{yy}(t,s) = \int_0^T \int_0^T g(\sigma,t)g(\tau,s)R_{xx}(\sigma,\tau)\,ds.$$

Упражнения

- Пусть $Y(t) = \alpha(t)X(t) + \beta(t)X'(t)$. Тогда

$$R_{yy}(s,t) = \alpha(t)\alpha(s)R_{xx}(s,t) + \alpha(t)\beta(s)\frac{\partial R_{xx}(s,t)}{\partial s} + \\ + \alpha(s)\beta(t)\frac{\partial R_{xx}(s,t)}{\partial t} + \beta(t)\beta(s)\frac{\partial^2 R_{xx}(s,t)}{\partial s \partial t}. \quad (?)$$

- Если случайный сигнал $X(t)$ имеет корреляционную функцию

$$R_{xx}(s,t) = e^{-|t-s|},$$

[7] В том числе — *функцией Грина*, см. [5, т. 2].

то корреляционная функция интеграла $Y(t) = \int_0^t X(s)\,ds$ равна

$$R_{yy}(s,t) = 2\min\{s,t\} + e^{-s} + e^{-t} + e^{-|t-s|} - 1. \quad (?)$$

6.7. Системы регулирования

Понимание метаморфоз, которые происходят со случайными сигналами при интегрировании и дифференцировании, играет важную роль в изучении динамических систем типа

$$\ddot{X} + \beta(t)\dot{X} + \gamma(t)X = Y(t),$$

где параметры и внешние силы флуктуируют случайным образом.

В теории автоматического регулирования, например, рассматривается модель

$$Lx = My,$$

в которой L и M — дифференциальные операторы, y — вход системы, x — выход.

Преобразование Лапласа $Lx = My$ дает[8)]

$$L(p)\widehat{x}(p) = M(p)\widehat{y}(p),$$

$L(p)$ и $M(p)$ — обычные характеристические полиномы.

В результате

$$\widehat{x}(p) = W(p)\widehat{y}(p),$$

где

$$\boxed{W(p) = \frac{M(p)}{L(p)}}$$

называют *передаточной функцией* системы.

При работе с устойчивыми системами[9)] в качестве передаточной используется функция $W(i\omega)$, которая по Фурье-преобразованию входного сигнала $\widehat{y}(i\omega)$ позволяет указать Фурье-преобразование выходного сигнала $\widehat{x}(i\omega) = W(i\omega)\widehat{y}(i\omega)$.

[8)] Например, $Lx = \ddot{x} + T_1\dot{x} + T_0x \;\Rightarrow\; L(p) = p^2 + T_1p + T_0$.

[9)] Когда борьба с расходящимися интегралами не подталкивает к использованию преобразования Лапласа вместо — Фурье.

В отличие от детерминированных систем, преобразование Фурье выходного сигнала, рáвно как и сам сигнал, — для понимания ситуации ничего особенно не дают. Здесь важны не беспорядочные флуктуации, а вероятностные характеристики сигнала, определяемые преобразованием спектра:

$$\widehat{R}_{xx}(\omega) = W(i\omega)W(-i\omega)\widehat{R}_{yy}(\omega),$$

т. е.

$$\boxed{\widehat{R}_{xx}(\omega) = |W(i\omega)|^2 \widehat{R}_{yy}(\omega).} \tag{6.9}$$

Другими словами, при прохождении случайных сигналов через линейные системы основную роль играет не сама передаточная функция $W(i\omega)$, а ее модуль $|W(i\omega)|$.

В результате простого вычисления (6.9) по спектру входного случайного сигнала определяется спектр выходного сигнала. Для вероятностного анализа это практически вся информация, которая требуется.

6.8. Задачи и дополнения

- Эргодичность с. ф по отношению к корреляционной функции обеспечивается условием

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \left(1 - \frac{\sigma}{T}\right) [R_{xx}^2(\sigma) + R_{xx}(\sigma+\tau)R_{xx}(\sigma-\tau)]\, d\sigma = 0$$

при любом τ, а эргодичность по отношению к дисперсии — условием

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \left(1 - \frac{\sigma}{T}\right) R_{xx}^2(\sigma)\, d\sigma = 0.$$

- Если τ_t обозначает время в промежутке $[0, t]$, проводимое броуновской частицей на положительной полуоси, то, как обнаружил П. Леви,

$$\mathbf{P}\{\tau_t < xt\} = \frac{2}{\pi} \arcsin \sqrt{x},$$

что называют *распределением арксинуса*. По поводу дискретных аналогов см. раздел 7.3, а также [23, 26].

Глава 7

Прикладные области

Многие из упоминавшихся ранее задач являются прикладными, но некоторые из них формируются в направления. Об этом, собственно, идет речь. Не с целью достичь горизонтов, а с намерением дать представление.

7.1. Управление запасами

Продажа скоропортящегося товара сопровождается постоянным стрессом. Перезаказал — выбросил, недозаказал — упустил прибыль. Вся жизнь, конечно, такая. Но в продуктовом магазине — особенно.

Допустим, торговое время разбито на периоды, с. в. X обозначает спрос на товар внутри периода, x — объем заказа. Непроданный товар в течение «периода» приходит в негодность. Далее, λ — розничная цена, μ — оптовая. Прибыль, без учета привходящих факторов (накладные расходы, транспорт и т. д.), равна

$$\Pi(x, X) = \begin{cases} (\lambda - \mu)X - \mu(x - X), & X \leqslant x; \\ (\lambda - \mu)x, & X > x. \end{cases}$$

Пусть $F(x) = \mathbf{P}\{X < x\}$ — непрерывная функция распределения с. в. X, тогда

$$\frac{\partial \mathbf{E}\{\Pi(x, X)\}}{\partial x} = (\lambda - \mu)F(x) - \mu + \mu F(x).$$

Приравнивая эту производную нулю, получаем, что максимум условного матожидания прибыли достигается при заказе x, обеспечивающем равенство

$$F(x) = \frac{\mu}{\lambda}.$$

Для определения оптимального заказа, разумеется, необходимо знать $F(x)$ в диапазоне, который имеет отношение к реальной ситуации, — для чего требуется наблюдать и накапливать данные.

Постановка задачи, безусловно, игрушечная, но на готовый каркас легко нанизывать дополнительные детали. Кроме того, игрушечная модель выводит мысль из состояния замешательства и дает импульс в продуктивном направлении.

7.2. Страховое дело

Клиент страхует собственность на сумму X. Страховой взнос γX, вероятность потери собственности p.

Матожидание суммы потерь равно pX, — поэтому страховая компания будет «в плюсе» лишь при условии $\gamma X > pX$, т. е. $\boxed{\gamma > p}$. Использование *среднего* в данном случае логично, поскольку компания имеет дело с массой клиентов, — и картина в целом определяется действием закона больших чисел. В противном случае опора на матожидание была бы сомнительной.

Для индивидуального клиента картина совершенно иная. Массовость ситуации его не касается. Небеса подбрасывают «его монету» один раз — и усреднять нечего.

Это одно из противоречий бытия. Судьба армии мало что говорит об индивидуальном пути солдата. Матожидание не гарантирует отдельных результатов, и выигрыш в среднем иногда равносилен проигрышу наверняка, о чем уже не раз заходила речь (см. раздел 4.1). Поэтому в такого рода ситуациях оптимизация матожидания далека от реальных потребностей. Адекватные постановки задачи возможны лишь на базе содержательного понимания проблемы. Итогом может быть, например, максимизация вероятностей тех или иных событий, которые определяются факторами, лежащими за пределами исходного описания.

Выход из тупика «если страхование выгодно для компании, то оно невыгодно для клиента» нередко преподносится как некое таинство, опирающееся на громоздкие формулы. На самом деле возможная целесообразность страхования для клиента, как правило, опирается на очень простое соображение: субъективная ценность страхуемой собственности может быть гораздо выше ее рыночной стоимости X.

Возьмем крайний случай. Пусть речь идет о страховании автомобиля, субъективная ценность которого с точки зрения владельца

может быть бесконечной в следующем смысле. Потеря автомашины (из-за отсутствия денег на покупку новой) может быть связана с потерей работы, расположения любимой девушки и т. п. Таким образом, «в случае чего» клиент теряет жизненно важные точки опоры. Поэтому для него целесообразна любая *посильная* плата γX за страховку.

В этом, собственно, и заключена суть — в расхождении стоимости и рыночной цены. Бывает, например, недвижимость ничего не дает кроме головной боли. И продать жалко, и толку — чуть. Стоимость меньше цены — страховка неразумна [1].

Довольно много моделей опираются на страхование части рыночной стоимости X и подсчете средних потерь (для клиента!) за несколько периодов. Там возникает много формул — и на гуманитариев это действует гипнотизирующе.

7.3. Закон арксинуса

Случайное блуждание или игра в «орлянку», связанные с изучением сумм

$$S_n = X_1 + \ldots + X_n,$$

где $X_k \in \{-1, 1\}$ («выиграл — проиграл» либо «вверх — вниз»), — имеет существенное каноническое значение, важное для понимания различных содержательных задач.

Значения $t = k$ удобно считать дискретными моментами времени, а поведение S_n представлять как график в плоскости (t, S), на котором точки (k, S_k) соединены прямолинейными отрезками.

7.3.1 Теорема. *Пусть $p(2n, 2k)$ обозначает вероятность того, что в интервале времени $(0, 2n)$ сумма S_j принимает неотрицательное значение (выигрыш $\geqslant 0$) при $2k$ значениях j. Тогда*

$$\boxed{p(2n, 2k) = 2^{-2n} C_{2k}^{k} C_{2n-2k}^{n-k}.} \tag{7.1}$$

[1] Криминальные варианты «застраховать и поджечь» здесь не обсуждаются.

Несложное, но несколько перегруженное деталями доказательство см. у Феллера [26]. В основе доказательства лежит механизм, который в наиболее прозрачном виде действует в классической *задаче о баллотировке*:

На выборах кандидат A собрал a голосов, кандидат B — b голосов $(a > b)$. Вероятность, что в течение всего времени A был впереди B, равна $\dfrac{a-b}{a+b}$.

Эту задачу обычно сопровождают различные трактовки, выводящие на довольно широкий спектр приложений.

Чтобы рассмотреть суть за фасадом формулы (7.1), надо перейти к асимптотике, что одновременно полезно с вычислительной точки зрения. Кроме того, интерес представляет доля времени, когда выигрыш неотрицателен, т. е. вероятность

$$\mathbf{P}\{k_n < xn\} = \sum_{k=1}^{k_n} p(2n, 2k),$$

где $2k_n$ равно числу значений $j \in (0, 2n)$, при которых сумма S_j неотрицательна.

После некоторой технической эквилибристики получается, что при $n \to \infty$

$$\boxed{\mathbf{P}\{k_n < xn\} \to \frac{2}{\pi} \arcsin \sqrt{x}.}$$

Это и есть *закон арксинуса*. Вот несколько «цитат» из Феллера [26], компенсирующих сухость формулы.

При 20 бросаниях симметричной монеты в «орлянке» один из игроков с вероятностью 0,35 никогда не будет впереди, и с вероятностью 0,54 — будет впереди не более одного раза.

Интуиция подсказывает, что доля времени k_n/n, когда суммы S_j неотрицательны, должна быть близка к $1/2$. Но это как раз наименее вероятно. Наибольшую вероятность имеют крайние значения $k_n/n = 0$ и $k_n/n = 1$.

Выглядит абсурдно, но тем не менее вероятность того, что при 10 000 бросаний монеты один из игроков находится в выигрыше более чем 9 930 раз, а другой — менее чем 70, больше 10 %.

Если говорить простым языком, то причина разобранного явления заключается в том, что суммы S_j обнуляются все реже и реже. Подсознательно думается, что число ничьих (обнулений S_j) пропорционально длине игры n. На самом деле их число пропорцио-

нально $\sqrt{n}$. Если построить график $S = S_t$, то это будет колебание со все увеличивающейся длиной волны и растущей амплитудой.

7.4. Задача о разорении

Игры, связанные с бросанием монеты, кажутся наивными, но в них играют все экономические субъекты. От крупных банков до физических лиц. Поэтому сопутствующая тематика важна не столько даже для максимизации прибыли, сколько для понимания окружающей среды и собственной роли в будничном коловращении.

Суть дела чаще всего проста, но некоторые явления имеют источником не вполне очевидные математические факты. Первое впечатление о тривиальности поведения случайных «01»-последовательностей не совсем верно. Среднее, конечно, — нуль, дисперсия — одна четвертая. Но даже «нормальная» асимптотика, позволяющая легко оценивать доверительные интервалы и другие нюансы, оставляет кое-что вне поля зрения.

Вопрос заключается в том, как ведут себя индивидуальные траектории

$$S_n = X_1 + \ldots + X_n, \quad X_k \in \{0, 1\}.$$

Дипломатичный ответ «когда как» не отражает всю правду. В поведении случайных сумм S_n есть общие закономерности. Некоторые естественные ожидания рушатся под давлением *закона арксинуса*, показывающего, что при игре в «орлянку» нет, например, никакой тенденции к выравниванию периодов лидерства.

Кроме того, закон больших чисел и предельные теоремы оставляют без внимания многие естественные вопросы, игровой аспект которых может быть первоочередным.

Вероятность разорения. *Допустим, при игре в «орлянку» ставка каждой партии равна* 1 *юаню, начальный капитал игрока* N *юаней. Игра прекращается в случае разорения (обнуления капитала) либо по достижению капиталом игрока величины* A. *Какова вероятность разорения* $p(N)$?

◀ Пусть событие R обозначает разорение игрока, V_+ — выигрыш в первой партии, V_- — проигрыш. Тогда

$$p(N) = \mathbf{P}\{R\} = \mathbf{P}\{R|V_+\}\mathbf{P}\{V_+\} + \mathbf{P}\{R|V_-\}\mathbf{P}\{V_-\},$$

и в силу

$$\mathrm{P}\{V_+\} = \mathrm{P}\{V_-\} = \frac{1}{2}, \quad \mathrm{P}\{R|V_+\} = p(N+1), \quad \mathrm{P}\{R|V_-\} = p(N-1),$$

получается

$$p(N) = \frac{1}{2}[p(N+1) + p(N-1)], \quad \text{т. е.} \quad p(N+1) = 2p(N) - p(N-1).$$

Решая последнее рекуррентное уравнение при очевидных краевых условиях $p(0) = 1$, $p(A) = 0$, приходим к

$$\boxed{p(N) = 1 - \frac{N}{A}.} \quad \blacktriangleright$$

Разумеется, вероятность достижения капиталом игрока суммы A равна

$$1 - p(N) = \frac{N}{A}.$$

Упражнения

- При игре в ту же игру, но с вероятностью выигрыша в каждой партии, равной $\nu < 1/2$, возникает рекуррентное уравнение

$$p(N) = \nu p(N+1) + (1-\nu)p(N-1),$$

решением которого служит

$$p(N) = \frac{\xi^A - \xi^N}{\xi^A - 1}, \quad (?) \tag{7.2}$$

где $\xi = (1-\nu)/\nu$.

- Анализ показывает, что вероятность разорения зависит также от величины ставки в отдельной партии. Извлекая на свет эту зависимость, можно получить ответы на некоторые неочевидные вопросы. Выгоднее ставить по юаню или по доллару? Сразу все или «по чуть-чуть»?
  Например, при игре в рулетку $p = 18/38 \approx 0{,}47$ — и единичная ставка в каждой отдельной партии — в соответствии с (7.2) — удваивает капитал $N = 20$ с вероятностью

$$1 - p(20) = \frac{1 - (20/18)^{20}}{1 - (20/18)^{40}} \approx 0{,}11.$$

  Вероятность же удвоения капитала при одноразовой ставке $N = 20$ в четыре раза больше, $p = 0{,}47$.

- Из предыдущего примера напрашивается вроде бы философский вывод: «чем меньше партий играешь в проигрышную игру[2)], тем лучше». Удивительно, но даже это не всегда так.

[2)] В данном случае игра проигрышна, поскольку $p = 0{,}47 < 0{,}5$.

Допустим, игрок выигрывает (проигрывает) серию из $2n > 0$ партий в рулетку, если его суммарный выигрыш больше ($\leqslant$) нуля. При наличии права выбора числа $2n$ (заранее) — вероятность выигрыша серии максимизирует $2n = 24$, а не $2n = 2$, как подсказывает внутренний голос[3].

- Если двое, A и B, с начальными капиталами a и b, играют в «орлянку», то средняя продолжительность игры[4] до разорения одного из игроков — равна ab. (?) Таким образом, если капитал первого — доллар, а второго — миллион, то ожидаемая продолжительность игры — миллион партий (хотя A, казалось бы, может очень быстро проиграть). Но здесь уместно вспомнить о ситуациях, когда $X_n \xrightarrow{P} 0$, но $\mathbf{E}\{X_n\} \to \infty$.

7.5. Игра на бирже и смешанные стратегии

Скрытность вероятностных процессов вкупе с человеческой страстью к загадочности покрывают биржевые игры мистическим туманом. Свой вклад вносит также тенденция профессионалов морочить головы заказчикам. Рекомендациям покупать убыточные акции даются такие заумные толкования, что клиенты платят за консультации, как загипнотизированные.

Суть дела, между тем, достаточно проста, но не настолько тривиальна, как в обыкновенных задачах оптимизации. Игрок на бирже действительно сталкивается с ситуациями, к которым человеческая психика не была подготовлена охотой на мамонтов.

В формализованном виде задача может выглядеть так. Игрок имеет n различных стратегий (покупки разных акций, например) в условиях неопределенности состояния экономической среды, характеризуемого m вариантами. Каждой комбинации возможностей отвечает свой выигрыш a_{ij}, где первый индекс указывает номер стратегии игрока, второй — состояния среды.

Таким образом, игрок в матрице выигрышей

$$A = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix}$$

выбирает строку, а «природа» (или злой рок) — столбец.

---

[3] См. *Dubins L., Savage L.* How to gamble if you must. New York: McGraw-Hill, 1965.

[4] В случае единичной ставки при каждом отдельном бросании.

Типичная игровая ситуация, когда нет возможности собственными действиями определить исход. Результат зависит еще от действий противной стороны — другого игрока или случая.

Вариантов понимания целесообразного принятия решения — есть много. Остановимся на главном и наиболее принципиальном.

Пусть, например,

$$A = \begin{bmatrix} 7 & 2 & 9 \\ 2 & 9 & 0 \\ 9 & 0 & 11 \end{bmatrix}.$$

Выбор первой строки (стратегии) гарантирует выигрыш не меньше 2. Удивительно, но если игра повторяется, то можно обеспечить средний выигрыш не меньше 5. Для этого надо применять, как говорят, *смешанную стратегию*: каждый раз выбирать не определенную, а i-ю строку с вероятностью p_i. Оптимальный в данном случае набор вероятностей:

$$\{p_1^*, p_2^*, p_3^*\} = \left\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right\}.$$

> *К идее вероятностных стратегий многие уже привыкли, но в принципе — это революционный шаг, прорыв в понимании окружающей действительности, обнаруживающий новые возможности.*

В общем виде ситуация выглядит следующим образом. Имеется два игрока. Первый — выбирает строки матрицы A с вероятностями $\{p_1, \ldots, p_n\}$, второй игрок — выбирает столбцы с вероятностями $\{q_1, \ldots, q_m\}$. Матожидание первого тогда равно

$$W(p, q) = \sum_{i,j} a_{ij} p_i q_j,$$

и он его, так или иначе, пытается максимизировать.

Если игра *антагонистическая* (выигрыш одного есть проигрыш другого), появляется естественная логика решения. Первый так выбирает $\{p_1, \ldots, p_n\}$, чтобы добиться максимума W при наихудшем для себя $\{q_1, \ldots, q_m\}$, второй — наоборот. В результате решением игры оказываются наборы вероятностей

$$p^* = \{p_1^*, \ldots, p_n^*\}, \quad q^* = \{q_1^*, \ldots, q_m^*\},$$

обеспечивающие равенство

$$\max_p \min_q \sum_{i,j} a_{ij} p_i q_j = \min_q \max_p \sum_{i,j} a_{ij} p_i q_j. \tag{7.3}$$

Решением (7.3) является седловая точка $W(p, q)$, которая всегда существует, что является теоремой, но это уже другая территория, и здесь нет резона останавливаться на доказательстве.

Если второй игрок — отклоняется от стратегии q^*, то выигрыш первого — возрастает. Следовательно, p^* гарантирует средний выигрыш не менее $W(p^*, q^*)$. Если второй игрок, однако, «изощрен, но не злонамерен», — как считал Эйнштейн, — то «логика седловой точки» становится менее убедительной, ибо можно добиться большего, располагая прогнозом действий противной стороны. Но это уже бесконечный путь оговорок и уточнений.

На фоне сказанного учет специфики именно биржевой игры достаточно очевиден, включая замену вероятностей пропорциями покупки различных акций. Реальность, конечно, намного сложнее рассмотренной модели. Но эта модель улавливает сам качественный механизм влияния неопределенности на целесообразность покупки или продажи акций. Разговоры о разумности приобретения убыточных акций — это всегда не вся правда. Покупка может быть целесообразной лишь в том случае, когда, пусть с малой вероятностью, но есть надежда на доход. Конкретно — надо взвешивать.

Еще один принципиальный момент — ориентация на средний выигрыш. Здесь опять надо учитывать возможность

$$\mathbf{E}\{X_n\} \to \infty, \quad \text{но} \quad X_n \xrightarrow{P} 0,$$

что уже не раз обсуждалось.

7.6. Процессы восстановления

Процессом восстановления называют параметрически заданную случайную величину

$$\eta(t) = \max\{k : S_k \leqslant t\},$$

где $S_k = \sum_{j=1}^{k} X_j$, а случайные величины X_j независимы и положительны.

Терминология проистекает из малопривлекательной, но удобной модели: X_1 — время исправной работы системы (прибора). После выхода системы из строя

она так или иначе *восстанавливается* (заменяется), время бесперебойной работы восстановленной системы — X_2, и так далее.

В таком сценарии S_k — время k-го *восстановления*, а $\eta(t)$ — *число восстановлений* до момента времени t.

Иногда под процессом восстановления подразумевают саму последовательность S_k, и тогда ясно, что можно говорить о своеобразном случайном блуждании «с переменным шагом все время вправо». Снятие ограничений соблазнительно, но оно ликвидирует всякую специфику, и задача растворяется в общем изучении сумм независимых с. в.

Тематика «восстановления» упоминается здесь с единственной целью. Это емкая и достаточно развитая область **ТВ**. Поэтому в случае возникновения определенного типа потребностей — полезно знать, что такая область есть, и знать ключевые слова, по которым можно найти зацепки. Само же «общевероятностное» образование вполне может обойтись без решения рутинных задач «восстановления», чтобы освободить голову для простых вещей. В общем курсе **ТВ** возможны, конечно, и другие акценты, если — не через край.

7.7. Стохастическое агрегирование

Упрощение задачи при увеличении размерности связано обычно с возможностью перехода на укрупненное описание системы. Классический образец такого перехода дает статистическая физика, но гипнотизирующая роль этого примера настолько велика, что за пределами термодинамики возможности агрегирования в значительной мере остаются скрытыми.

Допустим, имеется сложная сеть транспортных перевозок, или вычислительная сеть с многочисленными буферными устройствами, или система почтовой связи с большим количеством маршрутов и сортировочных узлов, или телефонная связь. Во всех этих случаях детальное описание функционирования объектов практически невозможно, но оно, безусловно, влияет на укрупненные показатели. Сколько требуется, например, автомобилей для удовлетворительной перевозки грузов по сети? Для точного ответа нужен подробный анализ: распределение автомобилей по маршрутам, расписание,

пропускная способность узлов и т. п. Но часто оказывается, что ответить можно приближенно, причем этот ответ довольно точен и практически не зависит от детальной информации. Для обоснования такой независимости в каждом отдельном случае, разумеется, необходимо самостоятельное исследование.

Можно ли, например, сказать «не глядя», чему равен максимальный поток в графе, если известна лишь суммарная пропускная способность дуг? Точнее, пусть граф Γ_n с n вершинами генерируется следующим образом. Дуга, соединяющая вершины i и j, появляется в Γ_n с вероятностью

$$p \in [\varepsilon, 1] \qquad (\varepsilon > 0),$$

а ее вес a_{ij} — реализация случайной величины, распределенной равномерно на $[0, 2]$.

Легко видеть, что среднее значение максимального потока $S(\Gamma_n)$ будет асимптотически стремиться к np, но само по себе это мало что дает. Аккуратные вычисления показывают, что

$$S(\Gamma_n)/np \xrightarrow{P} 1 \qquad (n \to \infty)$$

и это уже говорит о возможности в данном случае асимптотического агрегирования, т. е. о возможности игнорирования детальной информации.

Задачи подобного рода на регулярной основе почти не изучались, и здесь имеется обширное поле для исследований. Идеологическую опору асимптотического агрегирования по-прежнему составляют результаты типа нелинейного закона больших чисел, но это — подоплека. На поверхности лежат совсем другие вопросы. Например,

$$L = \frac{1}{n} \sum_k \min\{x_k, y_k\}, \tag{7.4}$$

причем известны только агрегаты $X = \sum_k x_k$ и $Y = \sum_k y_k$, по которым необходимо вычислять $L(X, Y)$.

Понятно, что до обоснования агрегирования — необходимо решить вопрос о подходящем выборе функции $L(X, Y)$, если таковая существует[5].

[5] Поиск вида $L(X, Y)$ в зависимости от тех или иных предположений о распределении x_k, y_k можно использовать в качестве упражнения.

Задача вида (7.4) может возникать в ситуациях следующего типа. В городе имеется n почтовых отделений, x_k обозначает поток за день клиентов в k-е отделение, y_k — пропускная способность k-го отделения. Тогда суммарный проходящий через отделения поток равен nL.

Специфика почтовой связи, разумеется, ни при чем. Пусть x_k — спрос на рынке k-го покупателя, y_k — предложение k-го продавца. Каждый покупатель случайно выбирает продавца, и тогда суммарный объем продаж оказывается равен nL. Конечно, механизм взаимодействия влияет на конечный результат, но в определенных предположениях существенны лишь среднестатистические характеристики этого механизма.

В некоторых прикладных задачах нелинейные функции большого числа переменных настолько завуалированы, что непосредственное применение результатов главы 3 выглядит проблематичным, и более естественными оказываются специфические пути.

При изучении, например, динамических систем большой размерности довольно естественной представляется попытка судить об устойчивости на основе неких усредненных показателей, характеризующих систему. С увеличением размерности подобные «макрокритерии» становятся все более надежными. Если система «устойчива в среднем» и размерность достаточно велика, то вероятность устойчивости системы сколь угодно близка к единице.

Уточнять сказанное можно различным образом. Пусть речь идет об асимптотической устойчивости равновесия дискретного процесса

$$x_{k+1} = Ax_k + b$$

или непрерывного

$$\frac{dx}{dt} = Ax - x + b,$$

где $A = [a_{ij}]$ — матрица размера $n \times n$; векторы $x, b \in R^n$.

Тогда при достаточно больших n и справедливости условия

$$\frac{1}{n}\sum_{i,j} |a_{ij}| \leqslant 1 - \varepsilon \qquad (\varepsilon > 0) \tag{7.5}$$

с большой вероятностью можно рассчитывать на положительный ответ в обоих случаях. Точнее говоря, если элементы a_{ij} равномерно

распределены на множестве (7.5), то вероятность того, что матрица $A - I$ гурвицева, — стремится к 1 при $n \to \infty$ [6].

Феномен стабилизации функций при больших размерностях довольно существенно влияет на понимание и трактовку задач оптимизации. Упрощенно говоря, суть дела заключается в следующем. Функция $\varphi(x_1, \ldots, x_N)$ «почти постоянна» — и окрестность максимума, в которой $\varphi(x)$ ощутимо превышает среднее значение, весьма мала. Поэтому ошибки исходных данных могут сводить на нет усилия, направленные на точное решение задачи. Задание, например, сотни параметров с округлением до второго знака, может в принципе менять значение целевой функции в несколько раз. Кроме того, из-за неточностей моделирования сами постановки оптимизационных задач при больших размерностях становятся непохожи на реальность. Возникает парадокс: чем больше учтено факторов, тем хуже становится модель. Тем не менее это обычно так. Использование небольшого числа переменных для описания системы свидетельствует, как правило, о том, что в задаче поймано и выделено главное, а большое число переменных — об обратном, о попытке вычислять температуру по движению отдельных молекул.

Поэтому постановки задач большой размерности нередко имеют условную ценность, и оптимизация там по сути осуществляется не с целью поиска наилучшего решения, а с тем, чтобы не попасть в чересчур невыгодный режим. В этом смысле теоремы о стабилизации дают необходимую гарантию, попасть в минимум — так же трудно, как и в максимум, и почти все выбранные наугад решения примерно одинаковы по качеству.

Это, конечно, декларация — для обострения разговора. Более логичны могут быть другие схемы. Скажем, выделение в

$$\varphi(x_1, \ldots, x_N) \to \max$$

двух-трех агрегатов с последующей заменой исходной максимизации — близкой задачей: $\psi(U, V) \to \max$. Либо замена детальных ограничений усредненными. Либо даже качественное переосмысливание исходной постановки.

Феномен стабилизации может играть определенную роль и в дискретной оптимизации. Там уже сложилась традиция для некоторых типов задач доказывать, что те или иные эвристические алгоритмы дают решение «не хуже среднего». Было бы полезно при этом еще доказывать, что почти все допустимые решения находятся в районе среднего. Это бы позволило в «непробиваемых» ситуациях ориентировать эвристику просто на поиск допустимого решения. Вот маленькая иллюстрация [7].

---

[6] *Опойцев В. И.* Устойчивые системы большой размерности // А и Т. 1986. № 6. С. 43–49.

[7] *Перепелица В. А.* Асимптотический подход к решению некоторых экстремальных задач на графах // Проблемы кибернетики. 1973. **26**. С. 291–314.

Дуга у n-вершинного графа Γ_n появляется с вероятностью

$$p_n \geqslant \sqrt{2\frac{\ln n}{n}},$$

ее длина r_{ij} — реализация случайной величины, равномерно распределенной на $[0, 2r]$. Тогда при достаточно больших n почти все графы Γ_n имеют хотя бы один гамильтонов контур и длина почти всех гамильтоновых контуров стабилизируется около nr.

В оптимизации может представлять интерес и другой «асимптотический ракурс». Дело в том, что эвристические алгоритмы, оцениваемые в категориях правдоподобия и здравого смысла, могут при увеличении размерности давать асимптотически точные решения.

Вот тривиальный пример, дающий представление, о чем речь. Пусть в классической задаче о ранце имеется n предметов, v_i — стоимость i-го предмета, w_i — его вес. Надо выбрать группу предметов с максимальной суммарной стоимостью при ограниченном суммарном весе, т. е. решить задачу

$$\sum_i v_i x_i \to \max, \qquad \sum_i w_i x_i \leqslant W,$$

где x_i может принимать значение 1 или 0 («брать» или «не брать»).

Эффективные способы точного решения задачи отсутствуют. Естественный, но не оптимальный, алгоритм решения состоит в том, чтобы упорядочить предметы по удельной стоимости $c_i = v_i/w_i$, а потом поочередно складывать их в трюм корабля, пока соблюдается ограничение по весу. Легко видеть, что такой алгоритм будет при $n \to \infty$ асимптотически оптимален, если все $w_i/W \to 0$.

7.8. Агрегирование и СМО

Системы массового обслуживания (СМО) заслуживают отдельного упоминания, ибо представляют собой обширную область, где идеология асимптотического агрегирования, по существу, занимает центральное место. В целом теория СМО развивается по аналогии со статистической физикой. В фокусе внимания находятся макропараметры типа средней длины очереди или среднего времени

обслуживания, — зависящие от архитектуры системы и микроскопической организации ее работы, касающейся, в основном, выбора дисциплины обслуживания заявок.

Ориентация на статистические методы здесь естественна и эффективна. Несколько странно лишь отсутствие «термодинамического противовеса», который в физических приложениях играет полезную роль, смещая фокус внимания в иную плоскость. Поэтому было бы разумно ожидать развития «термодинамики СМО», что изменило бы акценты и расширило охват задач. Большой процент приложений из предыдущего раздела, например, вполне бы мог быть отнесен к СМО, — разумеется, при направлении мысли в другое русло.

Тормозом на этом пути, безусловно, является большое разнообразие описаний СМО на микроуровне. В физике — легче. Уравнения Гамильтона или Шрёдингера — в некотором роде исчерпывают варианты микроповедения изучаемых систем. А в массовом обслуживании каждый новый диссертант придумывает свою дисциплину обслуживания, не считая мелких «винтиков», — и все приходится начинать сначала. Понятно, что в такой ситуации до «термодинамики» руки не доходят.

В данном контексте представляют интерес «термодинамические закономерности», нечувствительные к микрооописанию СМО. Например, *формула Литтла*:

$$N = \lambda T, \tag{7.6}$$

где λ — интенсивность потока заявок, N — среднее число заявок в системе (или в очереди), а T — среднее время пребывания заявки в системе (или в очереди), — *при установившемся стационарном режиме*.

Факт (7.6), вообще говоря, тривиален, но интересно, что он нередко преподносится как крупное достижение, потому что формула оказывается верной независимо от характеристик потока и обслуживания заявок. Причина «неожиданности», разумеется, в привычке. Чувствительность макропараметров СМО к деталям микрооописания настолько характерна, что ее отсутствие, конечно, воспринимается как открытие.

7.9. Принцип максимума энтропии

Переход на укрупненное описание системы почти всегда может опираться на *принцип максимума неопределенности*, который хорошо

известен в термодинамике, но диапазон его применимости гораздо шире. Рассмотрим простой пример.

Пусть в городе имеется n районов, L_i — число жителей i-го района, W_j — число работающих в j-м районе, x_{ij} — число живущих i-м районе и работающих в j-м.

Очевидно,

$$\sum_j x_{ij} = L_i, \qquad \sum_i x_{ij} = W_j. \tag{7.7}$$

Величины L_i, W_j известны, необходимо оценить пассажиропотоки x_{ij}.

Для определения n^2 неизвестных в системе (7.7) имеется всего лишь $2n$ уравнений. Ситуация представляется неопределенной. Тем не менее соображения о «случайном» характере происхождения величин x_{ij} позволяют решить задачу практически однозначно.

Для фиксированных значений x_{ij} число способов расселения жителей равно

$$S(X) = \frac{N!}{\prod\limits_{i,j} x_{ij}!},$$

где $N = \sum L_i = \sum W_j$.

С учетом формулы Стирлинга $\ln k! \sim k \ln k$,

$$\ln S(X) \sim N \ln N - \sum_{i,j} x_{ij} \ln x_{ij}.$$

Поэтому максимум $S(X)$ достигается на той же матрице

$$X^* = [x_{ij}^*],$$

что и максимум «энтропии» [8)]

$$H = -\sum_{i,j} x_{ij} \ln x_{ij}.$$

В то же время ясно, что если набору $\{x_{ij}^*\}$ отвечает максимальное число способов расселения $S(X^*)$, то это и есть наиболее вероятное решение задачи. А если в некоторой ε-окрестности

[8)] В таком ракурсе задача рассматривалась в книге А. Вильсона «Энтропийные методы моделирования сложных систем» (М.: Наука, 1978).

решения $\{x_{ij}^*\}$ сосредоточены почти все возможные способы расселения, то это уже будет решением — с вероятностью, близкой к 1 (и стремящейся к 1 при $N \to \infty$). Именно такова рассматриваемая ситуация.

До пояснения этого факта остановимся на самом решении задачи

$$-\sum_{i,j} x_{ij} \ln x_{ij} \to \max, \qquad \sum_j x_{ij} = L_i, \qquad \sum_i x_{ij} = W_j.$$

Метод множителей Лагранжа легко приводит к $x_{ij} = e^{-1-\lambda_i-\mu_j}$, откуда ясно, что все x_{ij} представимы в виде произведения $x_{ij} = u_i v_j$. Подстановка в ограничения дает систему уравнений

$$\sum_j u_i v_j = L_i, \qquad \sum_i u_i v_j = W_j,$$

решая которую, окончательно имеем

$$\boxed{x_{ij}^* = \frac{L_i W_j}{N},}$$

что можно интерпретировать как наличие у районов «потенциалов притяжения», $\frac{L_i}{\sqrt{N}}$ и $\frac{W_j}{\sqrt{N}}$, произведение которых дает пассажиропоток x_{ij}^*.

Пониманию свойств решения x_{ij}^* мог бы способствовать какой-нибудь вероятностный сценарий. Вот один из возможных вариантов. Каждого жителя охарактеризуем распределением p_{ij} по n^2 состояниям (i, j) (жить в i-м районе, работать — в j-м). Если на прицеле держится задача «разбросать» N жителей по этим n^2 состояниям так, чтобы по матожиданию было соответствие с макроограничениями (7.7), то p_{ij} должны удовлетворять системе

$$\sum_j p_{ij} = \frac{L_i}{N}, \qquad \sum_i p_{ij} = \frac{W_j}{N}.$$

А если добавить максимум неопределенности[9]:

$$-\sum_{i,j} p_{ij} \ln p_{ij} \to \max,$$

[9] См. следующую главу.

то по виду — получается та же задача, и те же окончательные формулы:

$$x_{ij}^* = Np_{ij}^*.$$

Но теперь $x_{ij} = \sum \xi_{ij}$, где случайные величины ξ_{ij} равны 1 с вероятностью p_{ij}, и 0 — с вероятностью $1 - p_{ij}$. Большое число слагаемых гарантирует, в силу закона больших чисел, концентрацию почти всех вариантов в районе матожидания $x_{ij}^* = Np_{ij}^*$ с убыванием флуктуаций пропорционально $1/\sqrt{N}$. Другими словами, в малой ε-окрестности максимума $S(X^*)$ оказываются сосредоточены почти все возможные способы расселения, о чем и шла речь выше.

7.10. Ветвящиеся процессы

Каноническая модель простейшего ветвящегося *процесса Гальтона—Ватсона* рассматривает частицу, производящую себе подобные в количестве k штук с вероятностью p_k. Предмет изучения — динамика X_n, где с. в. X_n — количество частиц в n-й момент времени.

Процессы такого рода довольно широко распространены в различных областях. Динамика численности нейтронов при делении урана, распространенность того или иного гена (наследственного признака), возникновение эпидемий и т. п.

Первоначальным источником интереса к модели была проблема вырождения фамилий — обнуления траектории $\{X_n\}$, начиная с некоторого n_0, если подразумевается, что каждый мужчина фамильного рода с вероятностью p_k имеет k сыновей [10].

Пусть $G(z)$ обозначает производящую функцию распределения $\{p_0, p_1, \ldots\}$, т. е.

$$G(z) = \sum_{k=0}^{\infty} p_k z^k, \tag{7.8}$$

а $\Pi_n(z) = \mathbf{E}\{z^{X_n}\}$ — п. ф. X_n.

В случае $X_n = k$ с. в. X_{n+1} — есть сумма k независимых с. в. с распределением $\{p_0, p_1, \ldots\}$, — поэтому

$$\mathbf{E}\{z^{X_{n+1}} | X_n\} = [G(z)]^{X_n},$$

[10] Как всегда, речь идет об определенной идеализации, предполагающей в данном случае «синхронизацию поколений» и другие нюансы.

что после усреднения по X_n приводит к итерационной процедуре

$$\Pi_{n+1}(z) = \Pi_n[G(z)], \tag{7.9}$$

описывающей динамику X_n в терминах производящих функций.

Решением (7.9) в случае $X_0 = 1$ служит $\Pi_n(z) = G^{(n)}(z)$, где $G^{(n)}(z)$ обозначает n-ю итерацию $G(z)$.

Дифференцируя (7.9) и полагая $z = 1$, имеем, в силу (2.15),

$$\mathbf{E}\{X_{n+1}\} = \nu \mathbf{E}\{X_n\}, \qquad \nu = G'(1) = \sum_{k=0}^{\infty} k p_k.$$

Так что сходимость процесса по матожиданиям определяет значение ν. При $\nu \leqslant 1$ процесс сходится, в случае $\nu > 1$ — расходится. Из того же рекуррентного соотношения (7.9) легко извлекаются более тонкие результаты о поведении случайной последовательности X_n — см. [28].

Вероятность вырождения. Интуитивно достаточно очевидно, что X_n может сходиться либо к нулю, либо к бесконечности, и не может оставаться ненулевой ограниченной с ненулевой вероятностью, — разумеется, это теорема. Причем даже при больших ν вероятность обнуления X_n строго положительна.

Последовательность

$$q_n = \mathbf{P}\{X_k = 0;\ k \geqslant n\} = \Pi_n(0),$$

очевидно, монотонно растет, и потому имеет предел, $q_n \to q$ при $n \to \infty$. Величина q естественно интерпретируется как вероятность вырождения X_n.

В случае исходного положения $X_0 = 1$

$$q_{n+1} = G^{(n+1)}(0) = G[G^{(n)}(0)] = G(q_n),$$

откуда ясно, что q является корнем уравнения

$$\boxed{z = G(z).} \tag{7.10}$$

Из записи (7.8) легко видеть, что функция $G(z)$ выпукла, и уравнение (7.10) имеет два корня: один в любом случае равен 1, другой $q \leqslant 1$. Если $\nu > 1$, то $q < 1$. Если $\nu \leqslant 1$, то $q = 1$, т. е. процесс вырождается почти наверное. Пробелы рассуждения легко восполняются.

7.11. Стохастическая аппроксимация

На практике широко распространены оптимизационные задачи вида

$$\mathbf{E}\{Q(c, x)\} \to \min_c, \tag{7.11}$$

где усреднение идет по x, а минимизация — по c.

Такого сорта проблемы возникают в ситуациях, когда по случайному сигналу x надо делать те или иные выводы $y = Q(c, x)$, настраивая модель $Q(c, x)$ (вектором c) оптимально в среднем.

Это может быть *задача идентификации*: u, v — случайные *вход и выход* объекта, требуется построить модель $y = F(c, u)$, оптимальную по критерию минимума среднеквадратической ошибки

$$\mathbf{E}_{u,v}\{[v - F(c, u)]^2\} \to \min_c.$$

Ту же абстрактную форму имеет *задача распознавания* (*классификации*). Допустим, модель $y = F(c, x)$ предсказывает, к какому классу принадлежит объект x. Скажем, «$y = 1$», если к первому, и «$y = -1$», если ко второму. Естественный критерий в данном случае — *минимум ошибки распознавания*,

$$\int \{\rho_1(x)\theta[-F(c, x)] + \rho_2(x)\theta[F(c, x)]\}\, dx \to \min_c,$$

где ρ_1, ρ_2 — плотности распределения объектов первого и второго класса, а θ — функция Хэвисайда, равная 1 при положительном аргументе и 0 — при отрицательном.

К подобному классу относятся также задачи фильтрации, прогноза, минимизации рисков, потерь и т. п.

При известных плотностях распределения после интегрирования от вероятностной природы рассматриваемых задач ничего не остается. Однако плотности часто неизвестны, а если известны, то либо интегрирование непосильно, либо после усреднения возникают такие «монстры», что приходится искать другой выход, который обычно находят в *процедурах адаптации* — подстройки параметров в процессе наблюдения за объектом.

Итерационные детерминированные процедуры

$$z_{k+1} = \varphi(z_k) \quad \text{или} \quad z_{k+1} = z_k + \nu_k \varphi(z_k) \tag{7.12}$$

сходятся, если существует «функция Ляпунова» $V(z)$, убывающая на траекториях (7.12).

Нечто подобное имеет место и для стохастических процессов вида

$$c_{k+1} = c_k + \gamma_k \varphi(c_k, x_k). \tag{7.13}$$

Сходимость (7.13) — по вероятности или почти наверное — к решению уравнения $\boxed{\mathbf{E}_x\{\varphi(c, X)\} = 0}$ обеспечивается при существовании аналога «функции Ляпунова» $V(c)$, которая убывает на траекториях (7.13), но убывания теперь достаточно всего лишь в среднем:

$$\mathbf{E}_{x_k}\{V(c_k + \gamma_k \varphi(c_k, x_k))\} < V(c_k).$$

При этом коэффициенты $\gamma_k > 0$, регулирующие величину шагов, полагаются удовлетворяющими условиям

$$\sum_k^\infty \gamma_k = \infty, \qquad \sum_k^\infty \gamma_k^2 < \infty. \tag{7.14}$$

Первое из условий (7.14) не дает процедуре (7.13) остановиться раньше времени, а второе — предотвращает уход в бесконечность на маловероятных траекториях [11].

Процедуры стохастической аппроксимации типа (7.13) принято называть *процедурами Роббинса—Монро* [12]. В случае

$$\varphi(c, x) = \nabla Q(c, x)$$

процедура (7.13) решает задачу (7.11), а $\mathbf{E}_x\{Q(c, x)\}$ служит «функцией Ляпунова».

[11] Подобные меры необходимы и в детерминированном случае (7.12).

[12] *Robbins H., Monro S.* A stochastic approximation method // Ann. Math. Stat. 1951. **22**. P. 400–407.

Глава 8

Теория информации

При изложении теории информации естественная попытка ограничиться одной канонической моделью многое оставляет за бортом. Возникает дилемма: либо не смотреть по сторонам, либо мириться с переплетением обстоятельств, соглашаясь на определенную неуклюжесть.

8.1. Энтропия

Энтропия, как мера неопределенности, вещь довольно простая. Но как ахиллесова пята абстрактного мышления, она мистифицирует род людской не хуже Гарри Гудини, благодаря чему служит хорошей отдушиной для философских страстей.

В то же время надо признать, что налет загадочности у энтропии имеет основания. Термодинамическая сущность, не данная в ощущениях[1], — как говорится, не фунт изюму. Что касается сугубо информационного аспекта энтропии, то здесь, помимо неосведомленности о дробях и логарифмах, большую роль играет впечатление, что «$H = -\sum p_k \ln p_k$» обеспечивает вход в виртуальный мир, подтверждая его реальность. Однако — обо всем по порядку.

Неопределенность (*энтропия*) H при бросании m-гранной кости характеризуется наличием m возможностей. Интуитивно хотелось бы, чтобы при бросании двух костей[2] неопределенность была вдвое больше, т. е.

$$H(m^2) = 2H(m)$$

либо $H(mn) = H(m) + H(n)$, если кости имеют разное число граней.

---

[1] В отличие от температуры и давления.

[2] При котором число возможностей равно m^2.

Ясно, что такие предположения ведут к

$$\boxed{H(m) = K \ln m,}$$

что можно интерпретировать как $H(p, \ldots, p) = K \ln m$ при m равновероятных исходах, $p = 1/m$.

Следующий вопрос, как определить $H(p_1, \ldots, p_n)$ в случае *не* равновероятных исходов. Будем отталкиваться пока от следующей модели. Имеется несколько m_i-гранных костей. Число всевозможных граней равно $\sum m_i$, поэтому $H = K \ln \left(\sum m_i\right)$. С другой стороны, выбор может быть осуществлен в два приема. Сначала выбирается кость — ясно, что вероятности выбора числа граней при этом равны $p_i = m_i \Big/ \sum m_j$, — затем грань. Неопределенность первого шага — $H(p_1, \ldots, p_n)$, второго — средневзвешенная энтропия[3] $K \sum p_i \ln m_i$. Если потребовать *аддитивность*, т. е.

$$K \ln \left(\sum m_i\right) = H(p_1, \ldots, p_n) + K \sum p_i \ln m_i,$$

то

$$H(p_1, \ldots, p_n) = K\left\{\ln \left(\sum m_i\right) - \sum p_i \ln m_i\right\} =$$

$$= -K \sum \frac{p_i \ln m_i}{\sum m_j} = -K \sum p_i \ln p_i,$$

что *при непрерывной зависимости* H от аргументов будет справедливо и для иррациональных p_i.

От выбора константы K зависит лишь единица измерения энтропии. В случае $K = \dfrac{1}{\ln 2}$

$$\boxed{\boxed{H(p_1, \ldots, p_n) = -\sum p_i \log_2 p_i.}} \tag{8.1}$$

Здесь и далее действует соглашение $\boxed{0 \cdot \log 0 = 0}$. Двойка в основании логарифмов обычно опускается, а единица измерения называется *битом*. Таким образом, бит соответствует неопределенности выбора из двух равновероятных возможностей (то ли нуль, то ли единица).

(!) Комментарий. *Вернемся к использованию средневзвешенной энтропии*

$$K \sum p_i \ln m_i$$

[3] Потому что выбор на втором шаге зависит от реализации — первого. См. далее «Комментарий».

в описанной выше модели. Если выбор кости на первом шаге уже состоялся, — выбрана, скажем, 7-я кость, и речь идет об одноразовом опыте, — то неопределенность второго шага равна $K \ln m_7$, *и задача вырождается.*

О неопределенности обоих шагов естественно говорить в двух случаях:

- *либо задача решается до проведения опыта с оценкой того, что получится в среднем*[4];
- *либо опыт двукратного выбора повторяется много раз, и тогда матожидания типа* $K \cdot \sum p_i \ln m_i$ *возникают из-за частотной устойчивости эксперимента.*

Образно характеризуя ситуацию, можно сказать так. Гроссмейстер лучше новичка в шахматах понимает, что такое конь. Оба одинаково знают «как ходит», но у первого это вызывает ассоциации, у второго — ощущение дискомфорта. С энтропией такая же история.

Приведенный вывод (8.1) в рафинированном виде воспроизводит рассуждения Шеннона [30], наиболее просто выражающие суть дела. Но при первом знакомстве все же чувствуется определенная натяжка, избавиться от которой можно лишь расширив базу исходных примеров и ситуаций. Это — если говорить об индуктивном подходе. В другом варианте (8.1) принимается за определение энтропии, постулируются некие дополнительные свойства, — но далее все равно надо смотреть на примерах, как это работает.

Само по себе определение (8.1) мало что дает, поскольку при столкновении с действительностью возникает масса вопросов, не попавших в кадр. Положение облегчает следующая формальная схема, которая, если вдуматься, ничего принципиально нового не добавляет к бросанию костей, но все-таки увеличивает угол обзора.

Пусть $\{x_1, \ldots, x_n\}$ и $\{y_1, \ldots, y_n\}$ — возможные состояния случайных величин X и Y. Состояния вектора $\{X, Y\}$ представляют собой комбинации пар x_i и y_i. Энтропия $\{X, Y\}$ по определению равна

$$H(X, Y) = -\sum_{i,j} p_{ij} \ln p_{ij},$$

где $p_{ij} = p(x_i, y_j) = \mathrm{P}\{X = x_i,\ Y = y_j\}$.

[4] Собственно, другого варианта даже нет, поскольку исходы выбора могут ветвиться.

В описанной ситуации часто говорят, что имеется две системы X и Y с возможными состояниями $\{x_1, \ldots, x_n\}$ и $\{y_1, \ldots, y_n\}$. По существу ничего не меняется, но терминология иногда подталкивает мысль в новых направлениях.

Если системы X и Y независимы, то $p_{ij} = p_i p_j$ и

$$H(X, Y) = H(X) + H(Y), \tag{8.2}$$

что элементарно проверяется.

Если же системы зависимы, то $p(x_i, y_j) = p(x_i)p(y_j|x_i)$ [5], и

$$H(X, Y) = H(X) + H(Y|X), \tag{8.3}$$

где

$$H(Y|X) = \sum_i p(x_i)H(Y|x_i)$$

называют *полной условной энтропией*, а

$$H(Y|x_i) = -\sum_j p(y_j|x_i) \log_2 p(y_j|x_i)$$

условной энтропией Y при условии $X = x_i$.

В обоих случаях, (8.2) и (8.3), говорят об *аддитивности энтропии*. При независимости подсистем $H(Y|X) = H(Y)$ и (8.3) переходит в (8.2).

Некоторая чехарда при использовании (полной/неполной) условной энтропии связана с теми же обстоятельствами, которые обсуждались в «Комментарии» выше.

8.2. Простейшие свойства

- *Энтропия всегда неотрицательна и достигает максимума в случае равновероятных возможностей.*

Заметим, что решение любой задачи вида

$$\sum_k \varphi(p_k) \to \max, \qquad \sum_k p_k = 1$$

определяется решением системы уравнений

$$\varphi'(p_k) = \lambda, \quad k = 1, \ldots, n,$$

[5] Имеется в виду $p(x_i) = \mathbf{P}(X = x_i)$, $p(y_j|x_i) = \mathbf{P}(Y = y_j|X = x_i)$.

откуда ясно, что максимум достигается при равновероятных возможностях в очень свободных предположениях — не только в случае $\varphi(p) = p \log p$.

8.2.1 Лемма. *Пусть* $\sum p_k = \sum q_k = 1$, *т. е.* p_k *и* q_k — *два распределения, причем все* $q_k > 0$. *Тогда*[6)]

$$\boxed{\sum_k p_k \ln p_k \geqslant \sum_k p_k \ln q_k.} \tag{8.4}$$

◀ Введем в рассмотрение случайную величину X, принимающую значения q_k/p_k с вероятностями p_k. Очевидно, $\mathrm{E}\, X = \sum p_k \dfrac{q_k}{p_k} = 1$. Применяя к с. ф. $\ln X$ неравенство Иенсена (1.19), получаем (8.4). ▶

- *Условная энтропия всегда меньше или равна безусловной*

$$H(Y|X) \leqslant H(Y),$$

причем при добавлении условий энтропия не увеличивается.

◀ Лемма 8.2.1 гарантирует

$$\sum_j p(y_j|x_i) \ln p(y_j|x_i) \geqslant \sum_j p(y_j|x_i) \ln p(y_j).$$

Матожидание этого неравенства по X дает

$$\sum_{i,j} p(x_i, y_j) \ln p(y_j|x_i) \geqslant \sum_j p(y_j) \ln p(y_j),$$

что означает $H(Y|X) \leqslant H(Y)$ (знак минус перед суммами переворачивает неравенство).

Аналогично устанавливается справедливость оговорки об убывании энтропии при добавлении условий. ▶

8.3. Информационная точка зрения

Пусть $H(A)$ — энтропия исхода некоторого опыта A. Если опыт B содержит какие-то сведения относительно A, то после проведения B неопределенность A уменьшается до условной энтропии

[6)] Понятно, что в (8.4) ln можно заменить логарифмами по любому другому основанию.

$H(A|B)$. Разность

$$\boxed{I(A,B) = H(A) - H(A|B),}$$

по определению, есть *количество информации*, содержащееся в B относительно A. Равенство

$$\boxed{I(A,B) = I(B,A)}$$

вытекает из симметрии предполагаемого свойства (8.3).

Энтропия источника. На «микроуровне» это выглядит так. Если источник информации потенциально может передать i-й символ (алфавита) с вероятностью p_i, то величину информации при поступлении этого символа естественно принять за $-\log_2 p_i$. Математическое ожидание информации, либо ее среднее значение (на один символ) при длительной работе источника, будет равно

$$I = -\sum p_i \log_2 p_i,$$

т. е. — энтропии источника.

Здесь имеет смысл продумать старую схему в новых терминах. Если источник сообщает один из n равновероятных символов, то ... $I = K \ln n$, и далее — по уже готовой колее.

В итоге становится ясно, что информация и энтропия — это две стороны одного явления. Сколько поступает информации — настолько убывает энтропия (неопределенность). Чем больше энтропия источника[7], тем больше информации при получении его сигналов. Источник, способный генерировать единственный сигнал, никакой информации не производит. Источник, передающий только два сигнала «нуль/один», имеет единичную интенсивность (один бит на сигнал). Но при большой частоте способен производить много бит в единицу времени.

Пропускная способность канала. Канал связи в общей схеме

$$\boxed{\textit{источник}} \Rightarrow \boxed{\textit{канал связи}} \Rightarrow \boxed{\textit{приемник}},$$

[7] Об энтропии источника естественно говорить *до поступления* информации, *после* — логичнее говорить о производстве информации.

так или иначе, ограничивает скорость передачи информации. В простейшем и широко распространенном случае, когда символов (сигналов) всего два и их длительности одинаковы, *пропускная способность* C измеряется числом символов, способных пройти по каналу в единицу времени.

В общем случае C — это максимальная информация, которая может быть передана по каналу за одну секунду. Если, например, алфавит состоит из n букв и канал способен пропускать N букв в секунду (в точности или в среднем), то $C = N \log_2 n$.

Природа ограничений может быть различная. Скорость света, полоса пропускания частот, тактовая частота генератора[8]. Все это находится за рамками теории информации, но иногда понимание среды, в которой решаются задачи, играет важную роль.

8.4. Частотная интерпретация

Пусть источник генерирует i-й символ с вероятностью p_i, и символы в сообщении длины N независимы. При достаточно большом N количество символов i-го вида в сообщении с большой точностью равно Np_i. Это дает вероятность сообщения

$$p = p_1^{Np_1} \dots p_n^{Np_n},$$

т. е.

$$\log p = N \sum p_i \log_2 p_i \quad \Rightarrow \quad \boxed{p = 2^{-NH}}. \tag{8.5}$$

Иными словами, вероятности всех достаточно длинных сообщений равны $p = 2^{-NH}$, а поскольку эти сообщения еще и независимы, то их количество $K = 1/p$, т. е.

$$\boxed{\boxed{K = 2^{NH}.}} \tag{8.6}$$

Таким образом, энтропия по правилу (8.6) определяет, например, *количество текстов, в которых буквы встречаются с «правильной» частотой*. Если в определении энтропии вместо двоич-

[8] Упоминание в данном контексте генератора показывает, что ограничения *источника* могут быть «списаны» на ограничения канала связи.

ных используются натуральные логарифмы, то (8.6) заменяется на $K = e^{NH}$.

Если все p_i одинаковы, то $H = \log n$, и (8.6) приводит к максимально возможному числу сообщений: $K = n^N$.

Разумеется, количество текстов, в которых соблюдается заданная частотность букв, определяется формулой (8.6) с точностью до очевидных «ε-поправок». При чисто вероятностной (не частотной) трактовке требуются уточнения несколько иного рода. С какими бы вероятностями p_i источник ни генерировал символы — принципиально возможны все n^N сообщений Q длины N, но их вероятности $p(Q)$ различны.

Тогда при любом $\varepsilon > 0$

$$\lim_{N\to\infty} \sum_{|p(Q)-2^{-NH}|>\varepsilon} p(Q) = 0,$$

т. е. сумма вероятностей всех сообщений, вероятности которых отличаются от 2^{-NH} *более чем на* ε, *— стремится к нулю (сколь угодно мала при большом* N*).*

Соответственно, вероятности сообщений

$$p(Q) \in (2^{-NH} - \varepsilon,\ 2^{-NH} + \varepsilon)$$

в сумме стремятся к 1. Поэтому при больших N можно считать, что «наблюдаемых» сообщений (последовательностей, текстов) имеется как бы ровно 2^{NH}. Остальными можно пренебречь — их суммарная вероятность близка к нулю.

Описанная схема служит первым приближением к действительности, которым нередко и ограничиваются. Но более сложные методы вычисления энтропии заслуживают упоминания. Не столько по причине их практической значимости, сколько по теоретическим соображениям. Очевидно, например, что осмысленные тексты далеки от принятых выше предположений. Буквы в словах далеко не независимы — после гласной чаще следует согласная, а шестая буква шестибуквенного слова определяется по пяти предыдущим едва ли не однозначно.

Принципы определения энтропии в такого рода ситуациях идейно прозрачны. Допустим, имеет место «взаимодействие» соседних символов: j-й символ после i-го — может появиться с вероятностью p_{ij}. Энтропия *следующего* состояния в результате зависит от i и равна $H_i = -\sum_j p_{ij} \log_2 p_{ij}$. Если при этом P_i обозначают вероятности i-х состояний[9], то $H = \sum_i P_i H_i$.

[9] Стационарные вероятности марковского процесса.

8.5. Кодирование при отсутствии помех

Допустим, источник генерирует буквы из некоторого алфавита, и его энтропия равна H (*бит на символ*), а канал связи пропускает C (*бит в секунду*). Утверждать, что по каналу в среднем проходит C/H *символов в секунду*, конечно, нельзя — потому что результат зависит от качества кодирования. Но скорость $\boxed{C/H}$ асимптотически достижима [10] при оптимальном кодировании.

Если появление n символов (букв) *равновероятно*, то в секунду, очевидно, может проходить максимальное количество информации

$$I_{\max} = C \log n.$$

При использовании алфавита из двух символов $\{0, 1\}$, соответственно, $I_{\max} = C \log 2 = C$ *бит/с*.

Коэффициент избыточности сообщения определяется как

$$\frac{I_{\max} - I}{I_{\max}},$$

где I — количество информации в сообщении, а $I_{\max}$ — максимально возможное количество информации в сообщении той же длины.

Если символы *не равновероятны*, то на один символ в среднем приходится количество информации $-\sum p_i \log p_i < \log n$, и в результате $I < I_{\max}$. Подобное явление характерно для обычного текста — буквы (символы) появляются с различными частотами.

В то же время системы передачи информации, как правило, используют специальные символы, независимо от того, какого сорта информация передается (аудио, видео, текстовая). Общепринятый стандарт в цифровой технике «01»-последовательности.

Идея кодирования хорошо известна. Буквам, командам, операциям — сопоставляются различные последовательности вида 01...101. Иначе говоря, все описывается в *двоичном коде* — «01»-алфавите. В общем случае *кодирование* представляет собой запись исходной информации в любом другом алфавите по избранным правилам соответствия между группами символов.

[10] Кодирование способно обеспечить скорость $\geqslant C/H - \varepsilon$ при любом $\varepsilon > 0$.

Для конкретности, будем говорить о двоичном кодировании. Широко распространены: восьмибитовый [11] код EBCDIC [12] и семибитовый — ASCII [13]. Для русского текста семибитовой кодировки недостаточно — значительная часть двоичных комбинаций занята под латинские буквы и другие «надобности». Это было причиной появления восьмибитовой кодировки КОИ-8, а потом Windows-кода 1251.

Общепринято 8 бит (двоичных единиц) информации принимать за новую единицу измерения количества информации — один *байт*. Более крупная единица измерения — *килобайт* (1 Кбайт = 2^{10} байт = 1024 байта) [14].

Оптимальное кодирование. Одно и то же сообщение можно закодировать различным образом. Поэтому возникает вопрос о наиболее выгодном способе кодирования.

Естественное соображение: часто встречающимся символам и словам исходного сообщения ставить в соответствие короткие «01»-комбинации, редко встречающимся — длинные. Если удастся так закодировать сообщение, что символы 0 и 1 будут встречаться одинаково часто, — это будет оптимальным кодом.

Посмотрим, как это работает при кодировании русского алфавита. Среднестатистическая частота появления букв в текстах различна, — колеблется от $\sim 1/500$ для буквы «ф» до $\sim 1/10$ для буквы «о».

Оптимальную «игру» на длине кодовых комбинаций реализует *код Шеннона—Фано*. Буквы алфавита упорядочиваются по убыванию частоты (вероятности) p_i появления в тексте, после чего разбиваются на две группы. К первой — относят первые k букв — так, чтобы

$$\sum_{i=1}^{k} p_i \approx \sum_{i=k}^{n} p_i \approx \frac{1}{2},$$

после чего первой группе символов ставится в соответствие 0, второй — 1, и это определяет первый разряд кодового числа. Далее каждая группа снова делится на две приблизительно равновероятные подгруппы; первой подгруппе ставится в соответствие 0, второй — 1 и т. д. Группы с малым количеством букв быстро исчерпываются — и эти буквы в результате получают короткие коды. Легко убедиться, что в итоге кодовая запись достаточно длинного сообщения будет содержать приблизительно одинаковое количество нулей и единиц, т. е. при

[11] Буквы и команды кодируются восьмизначным двоичным числом — последовательностью из 8 символов 0 или 1.

[12] Аббревиатура от Extended Binary Coded Decimal Interchange Code.

[13] American Standards Committee for Information Interchange.

[14] Стандартная шутка: начинающий программист думает, что в килобайте 1000 байт, опытный — что в километре 1 024 метра.

любой частотности исходных символов частоты нулей и единиц двоичных кодов оказываются $\approx$ равны друг другу.

> *Обратим внимание, что изложение в главе, да и в книге, ведется в основном «с точностью до ε и других реверансов». За деталями можно обратиться к иным источникам, но гораздо важнее следовать иерархическим принципам изучения предмета, когда, скажем, идея предельного перехода не только перестает требовать расшифровки, но даже упоминания. В этом случае внимание не отвлекается на второстепенные подробности и концентрируется на главном.*

Информационная сторона оптимального кодирования очень проста, даже в самом общем виде. Вернемся к формуле (8.6). Равновероятные сообщения в количестве $K = 2^{NH}$ могут быть пронумерованы в двоичной записи, для чего потребуется минимальное число разрядов [15] $\log_2 K = NH$. Это и будет оптимальным двоичным кодом.

Минимум разрядов (символов в «01»-алфавите, электрических импульсов), необходимых для указания и передачи сообщения, означает наиболее эффективное использование канала связи (передачу максимума информации в единицу времени).

В рамках вероятностной модели возможны все n^N сообщений длины N (а не только $K = 2^{NH}$), но при больших N можно считать (см. предыдущий раздел), что «наблюдаемых» сообщений имеется как бы ровно 2^{NH}. Остальными можно пренебречь — их суммарная вероятность близка к нулю. Поэтому маловероятные сообщения можно кодировать достаточно длинными «01»-последовательностями. Из-за их маловероятности это в среднем почти не будет сказываться на скорости передачи информации.

> *Когда речь идет о минимуме числа разрядов в оптимальном коде, подразумевается, конечно, что алфавит задан. В алфавите из миллиона символов можно* ***одним символом*** *записать любое из миллиона сообщений. Но тогда надо иметь систему связи, способную генерировать и передавать миллион разных символов.*

[15] В m-ичной записи потребуется $\log_m K = NH$ разрядов.

Упражнения

- При энтропии источника H (*бит на букву*) и независимой генерации букв — оптимальное кодирование в среднем приводит к H двоичным знакам на букву. (**?**)
  Например, при бесхитростной нумерации букв русского алфавита в двоичной записи потребовалось бы 5 разрядов ($2^5 = 32$). С учетом частотности букв $H = -\sum p_i \log_2 p_i \approx 4{,}4$. Поэтому в среднем достаточно 4,4 знака на букву, что обеспечивает *код Шеннона—Фано*.
- В задачах оптимального кодирования чаще всего идет речь о перекодировании одних «01»-последовательностей в другие. Пусть энтропия источника «01»-сообщений равна H (*бит на символ*). Тогда длина n таких сообщений может быть уменьшена (за счет кодирования) до nH. (**?**)

8.6. Проблема нетривиальных кодов

Из предыдущего раздела следует, что при *оптимальном кодировании* необходимо отталкиваться от кодирования длинных сообщений. Не букв и даже не слов, а достаточно больших кусков текста. Тогда есть возможность достичь теоретического предела. Но технически удобнее, разумеется, *посимвольное* кодирование без дополнительных хлопот.

Поначалу кажется, что посимвольным кодированием можно обойтись, когда источник генерирует буквы независимо друг от друга. Это неверно.

Рассмотрим, например, источник, генерирующий две буквы, А — с вероятностью p, и Б — с вероятностью $1 - p$. Если p очень мало́, то любое посимвольное кодирование далеко от оптимального. Асимптотически оптимален RLE-код[16], суть которого состоит в сообщении длин серий[17] повторяющейся буквы Б.

Элементарные примеры типа RLE-кода создают иллюзию, что проблема кодирования тривиальна. На самом деле высокоэффективные коды являются часто результатом крупных достижений, с которыми все имеют дело, работая на компьютере, и не подозревая о наукоемкости различных архиваторов (ZIP, ARJ и др.). Элементом многих архивирующих программ является знамени-

[16] Аббревиатура от Run Length Encoding. Метод широко используется при передаче растровых изображений.

[17] Мы не вникаем в технические подробности кодирования, связанные, например, с синхронизацией, необходимой для отделения кодов одних символов от других.

тый *алгоритм Лемпеля—Зива*, осуществляющий *многоступенчатое кодирование*. Идея вчерне выглядит примерно так. Сообщение просматривается с помощью скользящего словаря, если в тексте появляется последовательность из двух ранее уже встречавшихся символов, то ей приписывается свой код, затем текст «прочесывается» на предмет повторяющихся комбинаций из большего количества символов, и так — до исчерпания текста.

Конечно, доведение идеи «до ума» сопряжено с преодолением массы сложностей, но здесь не место вдаваться в подробности, поскольку это территория другой научной дисциплины. Однако декорации при взгляде через призму теории информации играют вдохновляющую роль.

Очень интересны, например, методы MPEG (Moving Pictures Experts Group), которые при кодировании используют прогноз динамики изображений (передаются только меняющиеся пиксели). В результате достигается сжатие в несколько десятков раз.

Для сжатия данных неподвижных изображений широко используются методы JPEG (Joint Photographic Expert Group), исключающие малосущественную информацию (не различимые для глаза оттенки) за счет виртуозного использования преобразования Фурье.

Чтобы оценить возможные трудности оптимального кодирования, имеет смысл обратиться к простой на вид задаче о взвешивании монет (см. последний раздел главы), которая, по сути, есть задача оптимального кодирования [18]. Запутанность ее решения дает повод задуматься о *трудоемкости* кодирования, которая является существенным фактором, но остается за рамками информационного аспекта.

Оптимальный код — это совсем не то, к чему надо стремиться во что бы то ни стало [19]. Это лишь границы возможного, знание которых дает понимание ситуации.

8.7. Канал с шумом

При наличии шума в канале связи,

$$\boxed{\text{вход } X} \Rightarrow \overset{\Downarrow \xi}{\boxed{\text{канал связи}}} \Rightarrow \boxed{\text{выход } Y = f(X, \xi)},$$

[18] Сводящаяся к указанию номера фальшивой монеты в троичной записи.

[19] То же самое можно сказать о любых оптимизационных решениях.

выходной сигнал

$$Y = f(X, \xi)$$

зависит от входа X и шума ξ.

Если шум искажает в среднем 1 % символов, то о любом принятом символе нельзя сказать наверняка, правилен он или нет. Максимум возможного — при независимой генерации букв — утверждать их правильность с вероятностью 0,99. Но если речь идет о передаче осмысленного текста, то сообщение при 1 % ошибок можно восстановить (по словарю) с высокой степенью надежности. Понятно, что это возможно благодаря избыточности языка.

В общем случае проблема заключается в том, чтобы подобную избыточность использовать наиболее эффективно. Вернее даже — не использовать, а изобрести. Другими словами, *бороться с шумом специальным кодированием*. Разумеется, вероятность ошибки можно понизить за счет многократного повторения каждого символа, но это слишком неэкономно.

Для поиска рациональных путей необходимо понять сначала присущие задаче ограничения. Какова полезная информация, проходящая по шумящему каналу? Легко видеть, что это разность

$$I = H(X) - H(X|Y)$$

между уровнями неопределенности источника до и после приема сигнала Y. В нешумящем канале $H(X|Y) = 0$, т. е. принятый сигнал однозначно определяет переданный. В общем случае условная энтропия $H(X|Y)$ служит показателем того, насколько шумит канал.

При вероятности ошибки 0,01 в случае равновероятной передачи источником двоичных символов

$$H(X|Y) = -\frac{1}{100}\log\frac{1}{100} - \frac{99}{100}\log\frac{99}{100} \approx 0{,}08 \quad \textit{бит на символ.}$$

Поэтому при передаче по каналу 100 символов в секунду скорость передачи информации равна $100 - 8 = 92$ бита в секунду[20]. Ошибочно принимается лишь один бит из ста, но «потери» равны 8 битам из-за того, что неясно, какой символ принят неверно.

- Чему равна условная энтропия $H(X|Y)$ при том же уровне 0,01 ошибок, если источник генерирует 0 и 1 с вероятностями p и $1 - p$?
- В каких ситуациях $H(X|Y) = H(X)$?

[20] При $p = 1/2$, очевидно, $H(X|Y) = H(X)$, и скорость передачи информации нулевая, поскольку выходной сигнал не позволяет судить о входном.

Пропускная способность канала с шумом, по определению Шеннона, — это максимальная скорость прохождения информации

$$\boxed{C = \max[H(X) - H(X|Y)]} \quad (\textit{бит в секунду}),$$

где максимум берется по всем возможным источникам информации, а энтропия H измеряется в ***битах в секунду***.

На первый взгляд, это сильно отличается от канала без шума, где под C обычно мыслится максимально возможное число проходящих импульсов. Но это не совсем так. Во-первых, система передачи может быть не двоичной. Во-вторых, сама передача символов по каналу бывает малоэффективна — символов много, информации мало. Поэтому аккуратное определение пропускной способности канала без шума в точности совпадает с данным выше определением, при условии $H(X|Y) = 0$.

При этом ясно, что в ситуации $H > C$ передача информации без потерь невозможна [21]. В этом случае, кстати, на задачу можно смотреть как на передачу информации по специфически шумящему каналу.

В примере с искажением 1 % двоичных символов, если канал физически способен пропускать 100 бит/с, — его пропускная способность равна 92 бит/с. Информационные потери 8 бит приходятся на $H(X|Y)$, т. е. на шум.

Теоремы Шеннона. Допустим, что помимо основного — есть дополнительный корректирующий канал.

8.7.1. *Если корректирующий канал имеет пропускную способность не меньше $H(X|Y)$, то при надлежащей кодировке возможен практически безошибочный прием сообщений* [22] *(с точностью до сколь угодно малой доли ошибок).*

◀ На философском уровне утверждение самоочевидно. На приемном конце недостает $H(X|Y)$ бит/с информации — ее и надо передать по дополнительному каналу.

Если спуститься с небес на землю, то рассуждать можно так. Любому принятому сообщению достаточно большой длительности в t с — отвечает [23]

[21] Источник генерирует больше информации H (*бит в секунду*), чем пропускает канал.

[22] Имеется в виду, что информация $H(X) - H(X|Y)$ проходит по основному каналу.

[23] См. (8.6).

$K = 2^{tH(X|Y)}$ возможных равновероятных сообщений источника. Чтобы указать среди них правильное, нужна информация $tH(X|Y)$ бит, т. е. $H(X|Y)$ бит/с. ▶

Конечно, доказательство отдает метафизикой, но такова природа утверждения. Это теорема существования: *хорошо закодировать можно, но как — это уже другой вопрос*, не представляющий большого интереса (как показывает жизнь)[24].

8.7.2 Теорема. *Пусть H бит/с — энтропия источника, а C — пропускная способность канала с шумом. Если $H \leqslant C$, то при надлежащем кодировании возможен практически безошибочный прием сообщений (с точностью до сколь угодно малой доли ошибок).*

◀ Теорема 8.7.2 обычно позиционируется как в высшей степени интуитивно неожиданный результат. Однако неожиданность здесь проистекает из забывчивости интуиции, которая не помнит определения C в случае шумящего канала. На самом деле теорема 8.7.2 не что иное как переформулировка утверждения 8.7.1 при естественном допущении, что корректирующий канал с основным — могут быть объединены в один.

Посмотрим, что происходит в примере с искажением 1 % двоичных символов. Если канал физически способен пропускать 100 бит/с, — его пропускная способность равна 92 бит/с (см. выше). Тогда при $H \leqslant C$, т. е. при $H \leqslant 92$ бит/с остается 8 бит/с, которых как раз хватает для коррекции. ▶

Теорема 8.7.2 обычно дополняется утверждением, что *в случае $H > C$ по любому $\varepsilon > 0$ можно указать способ кодирования, при котором информационные потери будут не больше чем $H - C + \varepsilon$ бит/с.* В данном контексте — это легкое упражнение.

Коды Хэмминга. Жизнь обычно протекает вдали от фундаментальных ограничений типа абсолютного температурного нуля. Таковы же ограничения, устанавливаемые теоремами 8.7.1, 8.7.2. Реальное кодирование больше ориентируется на удобство и простоту. Широкое распространение получили несколько стандартных схем кодирования, в том числе кодирование по Хэммингу.

Расстояние по Хэммингу $h(A, B)$ между двоичными последовательностями одинаковой длины определяется как число разрядов, в которых A и B не совпадают. Например, $h(001, 100) = 2$.

[24] Оптимально кодировать обычно в голову не приходит, потому что достижение оптимума слишком трудоемко. Не говоря о том, что еще и декодировать приходится.

Если двоичные последовательности длины n интерпретировать как вершины куба n-мерного пространства, то $h(A, B)$ представляет собой минимальное число ребер, по которым можно перейти из A в B.

В случае, когда все расстояния между возможными сообщениями $h(A, B) \geqslant 2$, — любая одиночная ошибка (в двоичном разряде) будет обнаружена, а в случае $h(A, B) \geqslant 3$ — не только обнаружена, но и исправлена [25].

Идеологическая ясность не устраняет практическую задачу такого кодирования полезных сигналов, чтобы они были разнесены на заданное расстояние. «Зазор» $h(A, B) = 2$ легко обеспечивается введением дополнительного двоичного разряда, в который записывается 0 (или 1), в зависимости от четности (или нечетности) числа единиц в кодируемой двоичной последовательности. Бóльшие «зазоры» обеспечиваются иными ухищрениями, но это уже другая история.

8.8. Укрупнение состояний

Имея дело с тем или иным понятием, полезно располагать удобной для интуиции моделью. Что касается энтропии, то от содержательной интерпретации состояний системы всегда можно отвлечься и говорить только о номерах этих состояний, подразумевая случайную величину X, которая принимает некоторые значения, например, $X = k$ с вероятностями p_k.

Если состояния равновероятны, то $H = \log_2 n$ представляет собой количество двоичных разрядов, необходимых для записи всех чисел от 1 до n, а $H = \lg n$ — количество десятичных разрядов, необходимых для той же цели.

Если состояния не равновероятны, то

$$H = -\sum p_k \log_2 p_k < \log_2 n$$

равно *среднему* количеству двоичных разрядов, необходимых для записи чисел от 1 до n, но — возможно — при их *перенумерации* (*оптимальном кодировании*).

[25] Для исправления ошибочной последовательности $C = 0100\ldots10$ надо найти ближайшую к C разрешенную последовательность $A = 0101\ldots10$, которая, в силу одиночности ошибки, находится на расстоянии $h(A, C) = 1$.

Число состояний может быть даже бесконечно, рáвно как и число разрядов, необходимых для их записи. Но при условии $\sum_{k=1}^{\infty} p_k = 1$ среднее число разрядов будет равно как раз H.

Так или иначе, но для энтропии важны только вероятности состояний. Если с. в. X принимает значения 1 и 10 с вероятностями p и $1-p$, а с. в. Y с теми же вероятностями равна либо 1, либо $1+10^{-99}$, — то $H(X) = H(Y)$.

Другими словами, энтропия не ощущает неопределенности значений случайной величины. В то же время ясно, что «близкие» состояния системы иногда можно считать одинаковыми, объединяя их в одно состояние. Укрупнение возможно и по другим причинам. При этом энтропия $-\sum p_k \log_2 p_k$ переходит в

$$\widetilde{H} = -\sum_G p_G \log_2 p_G, \qquad p_G = \sum_{k\in G} p_k,$$

причем энтропия *укрупненной* (*агрегированной*) *системы* всегда меньше или равна исходной. (**?**) В случае разукрупнения системы энтропия, наоборот, увеличивается.

8.9. Энтропия непрерывных распределений

Энтропия случайной величины X, распределенной с плотностью $\rho(x)$, определяется как

$$H = -\int_{-\infty}^{\infty} \rho(x) \log \rho(x)\, dx. \tag{8.7}$$

Если X — случайный вектор, энтропия вычисляется по той же формуле с той лишь разницей, что интегрирование ведется по всему пространству.

Аналогия с дискретным случаем легко просматривается, но предельный переход к (8.7) невозможен, — по крайней мере, в общепринятом смысле.

Естественная аппроксимация (8.7) при разбиении оси x на промежутки Δx_k записывается в виде суммы

$$H_\Delta = -\sum_{k=-\infty}^{\infty} \rho(x_k)\Delta x_k \log \rho(x_k), \tag{8.8}$$

где x_k — некоторым образом выбранные точки на промежутках Δx_k. Функция $\rho(x)$ заменяется в результате ступенчатой аппроксимацией, а $p_k = \rho(x_k)\Delta x_k$ становится приближенной вероятностью попадания с. в. X на промежуток Δx_k. При этом (8.8) можно переписать в виде

$$H_\Delta = -\sum_{k=-\infty}^{\infty} p_k \log p_k + \sum_{k=-\infty}^{\infty} p_k \Delta x_k. \tag{8.9}$$

Фиксация $\Delta x_k = \varepsilon$ превращает второе слагаемое (8.9) в константу $c(\varepsilon)$. А поскольку не так важно, каков нулевой уровень неопределенности, то (8.7) *с разницей в константу* приближенно равно энтропии $-\sum p_k \log p_k$. Поэтому, если договориться, что энтропия измеряется с точностью, скажем, до третьего знака, то формулой (8.7) можно пользоваться как хорошим приближением (8.8).

Безболезненному оправданию предельного перехода мешает расходимость $c(\varepsilon) \to \infty$ при $\varepsilon \to 0$. Но из сказанного ясно, что большой беды в этом нет. Определение (8.7) вполне мотивированно, хотя и не совсем стандартным способом.

Свойства энтропии непрерывных распределений в основном аналогичны свойствам энтропии дискретных распределений. В частности, имеет место аддитивность вида (8.2) и (8.3) при естественной записи условной энтропии с помощью условной плотности, а также аналоги неравенств из раздела 8.2. Максимум энтропии на ограниченной области достигается при равномерной плотности. (?)

Максимум (8.7) при ограничениях

$$\int_{-\infty}^{\infty} \rho(x)\,dx = 1, \qquad \int_{-\infty}^{\infty} x^2 \rho(x)\,dx = \sigma^2$$

обеспечивает нормальный закон распределения [26)]

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}.$$

При этом $H(X) = \log \sqrt{2\pi e}\,\sigma$.

[26)] См. раздел 2.7.

Если случайные векторы X, Y функционально связаны линейным невырожденным преобразованием $Y = AX$, то

$$H(Y) = H(X) + \log \det A, \quad (?)$$

что легко проверяется, но заслуживает внимания, ибо здесь выявляются тонкости перехода к энтропии непрерывных распределений, о которых говорилось в начале раздела.

Наличие невырожденной функциональной связи $Y = AX$ в случае дискретного распределения к изменению энтропии не ведет, поскольку число состояний и их вероятности не меняются. В непрерывном случае аппроксимация (8.7) с помощью разбиения пространства на ячейки («промежутки» Δx_k) претерпевает изменения при линейном преобразовании переменных. Объемы ячеек, а значит, и соответствующие вероятности — меняются. Детерминант A дает как раз коэффициент искажения объема.

8.10. Передача непрерывных сигналов

Шеннон, создавший теорию информации — см. [30], начинает изучение непрерывных сигналов с *теоремы отсчетов*[27], которая сразу переводит задачу в плоскость дискретного времени.

Речь идет о следующем. Информационная емкость непрерывного сигнала $x(t)$ упирается в барьер точности. Важный ориентир в переплетении обстоятельств задает неизбежная[28] ограниченность спектра $x(t)$. В представлении Фурье[29]

$$x(t) = \int_{-\infty}^{\infty} \widehat{x}(\nu) e^{-2\pi i \nu t}\, d\nu \quad \Leftrightarrow \quad \widehat{x}(\nu) = \int_{-\infty}^{\infty} x(t) e^{2\pi i \nu t}\, dt$$

в условиях ограниченности спектра: $\widehat{x}(\nu) \neq 0$ только при $|\nu| < W$, — сигнал $x(t)$ представим в виде

$$x(t) = \int_{-W}^{W} \widehat{x}(\nu) e^{-2\pi i \nu t}\, d\nu. \tag{8.10}$$

Но $\widehat{x}(\nu)$, как функция, заданная на конечном промежутке $[-W, W]$, может быть разложена в *ряд Фурье* с периодом $2W$:

$$\widehat{x}(\nu) = \sum_{n=-\infty}^{\infty} a_n e^{in\pi\nu/W}, \tag{8.11}$$

---

[27] У нас ее принято называть *теоремой Котельникова* — см.: *Котельников В. А.* О пропускной способности «эфира» и проволоки в электросвязи // Материалы к I Всес. съезду по вопросам реконструкции дела связи. Изд. Упр-я связи РККА, 1933.

[28] Из-за конечности полосы пропускания частот любого канала связи.

[29] Обычно в преобразовании Фурье вместо частоты ν используется круговая частота $\omega = 2\pi\nu$, и тогда в первом интеграле появляется множитель $1/(2\pi)$.

где, с учетом (8.10),

$$a_n = \frac{1}{2W}\int_{-W}^{W} \widehat{x}(\nu)e^{-in\pi\nu/W}\,d\nu = \frac{1}{2W}x\left(\frac{n}{2W}\right). \tag{8.12}$$

Теперь подстановка (8.12) $\Rightarrow$ (8.11) $\Rightarrow$ (8.10) приводит к

$$x(t) = \frac{1}{2W}\sum_{n=-\infty}^{\infty} x\left(\frac{n}{2W}\right)\int_{-W}^{W} e^{\frac{i\pi\nu}{W}(n-2Wt)}\,d\nu,$$

что после несложных преобразований может быть переписано в виде

$$x(t) = \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2W}\right)\frac{\sin\pi(2Wt-n)}{\pi(2Wt-n)}. \tag{8.13}$$

Формула (8.13) показывает, что любой сигнал $x(t)$ с ограниченным спектром определяется значениями $x(t)$ в дискретном ряде точек, расположенных с интервалом времени $\Delta t = 1/(2W)$, который Шеннон называет *интервалом Найквиста*[30]. Факт может показаться удивительным, поскольку речь идет не о приближенном, а о точном воспроизведении сигнала по дискретным замерам. Но это удивление философского характера. На практике, понятно, вопрос точного воспроизведения никогда не стоит. В условиях ошибок измерения и других погрешностей говорить имеет смысл только об аппроксимациях $x(t)$, например кусочно-линейных, определяемых точно так же значениями сигнала в дискретном ряде точек. Особая роль соотношения (8.13) заключается в указании связи необходимого интервала замеров с шириной спектра сигнала[31].

В принципе, можно было бы ориентироваться на какую-нибудь аппроксимацию $x(t)$ типа *полиномов Бернштейна*,

$$P_n(t) = \sum_{k=0}^{n} x\left(\frac{k}{n}\right)C_n^k t^k(1-t)^{n-k},$$

равномерно аппроксимирующих $x(t)$ с любой наперед заданной точностью: $|x(t) - P_n(t)| < \varepsilon$. И тогда бы речь шла о передаче конечного числа коэффициентов $P_n(t)$, а теория — развивалась на прежней идеологической базе дискретных

[30] *Nyquist H.* Certain topics in telegraph transmission theory // AIEE Trans. Apr. 1928.

[31] Эта связь в какой-то степени метафизична, поскольку коренным образом зависит от требований к точности воспроизведения сигнала.

сообщений. Конечно, в поле зрения оказался бы включенным фактор точности, но в определенных условиях это было бы даже хорошо.

Вернемся, однако, к точке зрения Шеннона. Если функция $x(t)$ ограничена временны́м промежутком T, а замеры отстоят друг от друга на $1/(2W)$, то в промежутке T всего будет $\boxed{2TW}$ отсчетов[32)], которые всегда можно мыслить как координаты точки в пространстве $2TW$ измерений, причем из (8.13) легко следует

$$\int_0^T x^2(t)\,dt = \frac{1}{2W}\sum_{n=0}^{2TW} x^2\left(\frac{n}{2W}\right), \tag{8.14}$$

что в электросвязи, например, естественно интерпретируется как энергетическое соотношение.

Квадрат евклидова расстояния $\sum x^2\left(\frac{n}{2W}\right)$ оказывается равным $2WE$, где E — энергия, выделяемая на единичном сопротивлении при прохождении тока $x(t)$ на промежутке T. Поскольку $E = TP$, где $P = D_X$ — средняя мощность сигнала, то в силу (8.14) все сигналы с мощностью, меньшей P, будут расположены в шаре радиуса

$$r = \sqrt{2TWP}$$

либо $r = \sqrt{2WP}$, если рассматривать промежуток $T = 1$ с.

С точки зрения помехоустойчивости точки (сигналы) в этом шаре надо распределять равномерно, чтобы при заданном их количестве они были расположены как можно дальше друг от друга. Например, при аддитивной помехе:

$$Y(t) = X(t) + N(t),$$

где $X(t)$ — передаваемый сигнал, $Y(t)$ — принимаемый, $N(t)$ — белый шум мощности D_N. В силу независимости $X(t)$ и $N(t)$, мощность (дисперсия) сигнала на выходе равна

$$D_Y = D_X + D_N.$$

Объем «шумящего шарика», в силу $r = \sqrt{2TWD_N}$, оценивается

$$\sim \left(\sqrt{2TWD_N}\right)^{2TW},$$

[32)] Это очевидно даже без теоремы Котельникова. Найквист, например, рассуждал так. Разложение $x(t)$ в ряд Фурье на промежутке T содержит TW синусов и $(TW + 1)$ косинусов — вплоть до частоты W. Для определения $(2TW + 1)$ соответствующих коэффициентов достаточно $\approx 2TW$ замеров.

а объем шара выходных сигналов мощности $\leqslant D_Y$ —

$$\sim \left(\sqrt{2TW(D_X + D_N)}\right)^{2TW}.$$

Деление показывает, что маленьких шариков в большом помещается приблизительно:

$$\sim \left(\sqrt{\frac{D_X + D_N}{D_N}}\right)^{2TW},$$

т. е. в шар помещается приблизительно такое количество точек (сигналов), разнесенных на расстояние, не покрываемое шумом. Для записи этого количества требуется порядка

$$TW \log_2\left(1 + \frac{D_X}{D_N}\right) \quad \text{разрядов,}$$

что определяет число *бит/с*, которое можно передать по такому каналу за время T. При $T = 1$ с получается пропускная способность канала:

$$C = W \log_2\left(1 + \frac{D_X}{D_N}\right), \tag{8.15}$$

зависящая от полосы пропускания W и отношения *сигнал/шум*, D_X/D_N.

Несколько «лихой» вывод формулы Шеннона (8.15) имеет два оправдания. Во-первых, он в чистом виде отражает идею. Во-вторых, на точности соотношения (8.15) не имеет смысла особо настаивать, поскольку, строго говоря, здесь необходима масса оговорок. Но сам характер зависимости может служить путеводной нитью.

8.11. Оптимизация и термодинамика

При описании идеального газа (трехмерного бильярда) задания энергии не хватает для фиксации термодинамического состояния. Соображение максимизации неопределенности распределения скоростей молекул — решает проблему, определяя полный комплект макропараметров. И такое соображение работает во многих других ситуациях, где речь идет о статистическом описании сложных систем.

На формальном уровне это выглядит примерно так. Решается задача максимизации энтропии

$$-\sum_x \rho(x) \ln \rho(x) \to \max_{\rho(x)}$$

при ограничении [33]

$$\sum_x r(x)\rho(x) = R$$

и, разумеется, $\sum_x \rho(x) = 1$.

Стандартный переход к лагранжиану

$$L = -\sum_x \left[\rho(x)\ln\rho(x) + \lambda\rho(x) + \mu r(x)\rho(x)\right]$$

с последующим варьированием $\rho(x)$ в конечном итоге дает

$$\rho(x) = e^{-1-\lambda-\mu r(x)},$$

что с учетом нормировки, $\sum_x \rho(x) = 1$, приводит к

$$\rho(x) = \frac{e^{-\mu r(x)}}{\sum_x e^{-\mu r(x)}}, \tag{8.16}$$

где параметр μ определяется энергетическим ограничением.

Легко видеть, что вместо рассмотренной можно было бы решать другую задачу:

$$\sum_x r(x)\rho(x) \to \min_{\rho(x)}, \qquad -\sum_x \rho(x)\ln\rho(x) = H.$$

Ответ, с точностью до параметра μ («температуры» $T = 1/\mu$), был бы тот же самый, а при определенном соотношении R и H ответы бы совпали в точности.

Такая «взаимозаменяемость» задач широко используется в статистической физике, позволяя переходить, скажем, от максимизации энтропии к эквивалентной задаче минимизации энергии. Несмотря на физическую абсурдность второй задачи (энергия строго постоянна), ее рассмотрение математически оправдано и часто

[33] Строгое сохранение энергии обычно заменяется сохранением в среднем,

$$\sum_x r(x)\rho(x) = R,$$

что принципиально не меняет решения, но упрощает выкладки.

более удобно. Подобная «взаимозаменяемость» широко используется также в математическом программировании, где каждая задача, как правило, рассматривается в паре со своей двойственной.

В статистической физике для получения взаимосвязей между макропараметрами разработана удобная техника, опирающаяся на введение серии вспомогательных функционалов.

Сначала вводится статистическая сумма

$$Z = \sum_x e^{r(x)/T}$$

и свободная энергия $F(T) = -T \ln Z$, с помощью которой (8.16) записывается в виде (*распределение Гиббса*)

$$\rho(x) = e^{\frac{F - r(x)}{T}}.$$

Вычисление энтропии

$$H = -\sum_x \frac{F - r(x)}{T} e^{\frac{F-r(x)}{T}} = -\frac{F - r(x)}{T}$$

дает зависимость

$$F = R - TH.$$

Дифференцирование $F(T) = -T \ln Z$ по температуре,

$$\frac{dF}{dT} = -\ln Z - TZ^{-1}\frac{dZ}{dT} = -\ln Z - T^{-1}R,$$

с учетом $Z = \sum_x e^{r(x)/T}$ и $F = R - TH$ приводит к $H = -\dfrac{dF}{dT}$ и, в итоге,

$$R = F - T\frac{dF}{dT}.$$

А дифференциал

$$dR = d(F + TH) = dF + H\,dT + T\,dH,$$

в силу $H = -dF/dT$, оказывается равным $dR = T\,dH$, т. е.

$$\boxed{\frac{dR}{dH} = T,}$$

откуда следует, что «энергия» при возрастании энтропии H возрастает, если $T > 0$, и убывает, если $T < 0$.

8.12. Задачи и дополнения

- При наличии функциональной связи $Z = f(X, Y)$ величины X и Y дают полную информацию о Z. Возможно ли, что X и Y по отдельности не дают никакой информации о Z? Вопреки естественному ожиданию — возможно. Пусть X, Y, Z представляют собой n-разрядные числа в 10-тичной системе. Тогда число (функция) Z, определяемое поразрядным сложением по модулю 10,

$$Z_k = X_k + Y_k (\mathrm{mod}\ 10)$$

обладает нужными свойствами.
Например,

$$X = 123,\ Y = 948 \quad \Rightarrow \quad Z = 061.$$

Понятно, что задание X никак не уменьшает число возможных вариантов Z.

- Если все числа равновероятны, то деление группы n подряд идущих чисел на две равные подгруппы с последующим выделением одной из подгрупп (приписыванием, например, нуля или единицы) дает информацию $\log 2 = 1$, уменьшая исходную неопределенность $\log n$ до $\log n - \log 2$. После k аналогичных шагов неопределенность уменьшится до $\log n - k \log 2$ и станет $\leqslant 0$ при условии

$$k \geqslant \frac{\log n}{\log 2} = \log n.$$

Вот, собственно, и вся премудрость. Некоторые детали приходится уточнить, если n не является степенью двойки. Тогда группы чисел не делятся ровно пополам, и это уменьшает информацию некоторых шагов. Но легко проверить, что итог не меняется — из-за того, что $k \geqslant \log n$ выбирается целое. Задача становится совсем прозрачной при увеличении с самого начала n до ближайшего числа вида 2^m.
За кадром описанной схемы могут стоять разные интерпретации. От решения проблемы о числе вопросов при ответах «да — нет», необходимых для определения загаданного числа, — до указания числа разрядов для записи номера любого из n чисел в двоичной системе[34]. Двоичная запись чисел в последнем случае и будет оптимальным кодированием.

- Того же поля ягода простейшая задача о взвешивании монет. *Среди n монет есть одна фальшивая, более легкая. Найти минимальное число взвешиваний на чашечных весах*[35], *необходимое для определения фальшивой монеты в самом неблагоприятном случае.*

Любая из монет может равновероятно оказаться фальшивой, поэтому неопределенность равна $\log n$. Пусть пока $n = 3^m$. Разобьем монеты на три равные кучки, и любые две из них сравним по весу. Взвешивание (опыт B_1) может иметь три очевидных исхода. Любой — позволяет исключить две

[34] Либо самих чисел, если это числа от 1 до n.

[35] Позволяющих сравнивать два веса.

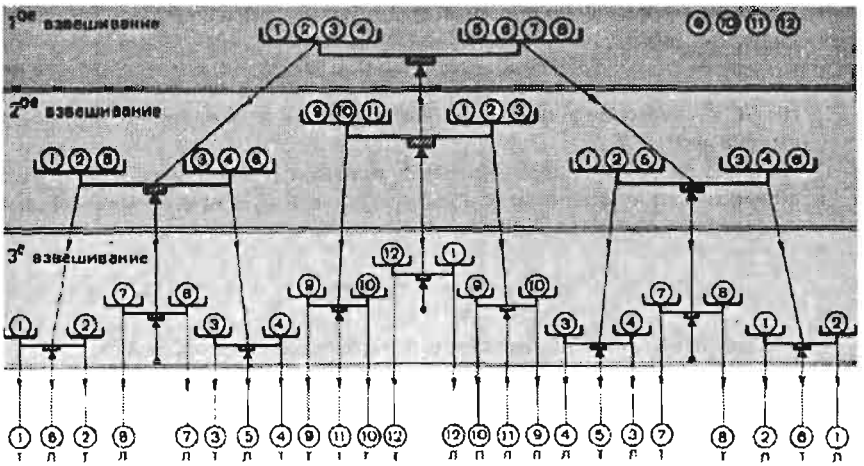


Рис. 8.1

группы монет. Неопределенность $H(B_1) = \log 3$. Энтропия (информация) k последовательных взвешиваний равна $k \log 3$. Для исчерпания исходной неопределенности $\log n$ необходимо $k \log 3 \geqslant \log n$, откуда $k \geqslant \log_3 n$. Легко убедиться, что ответ остается верным и в том случае, когда n не является степенью тройки.

- Если в предыдущей задаче неизвестно, легче или тяжелее фальшивая монета, то исходная неопределенность возрастает до $\log 2n$. Но естественный ответ $k \geqslant \log_3 2n$ уже не верен. Правильный ответ

$$k \geqslant \log_3(2n + 3),$$

однако это довольно сложная задача, что свидетельствует о трудностях оптимального кодирования[36]. Рецепт взвешиваний (кодирование) для 12 монет изображен[37] на рис. 8.1. Левая (правая) стрелка обозначает ситуацию, когда перетянула левая (правая) чаша весов, средняя стрелка отвечает равновесию. Все монеты перенумерованы, буквы Л, Т означают: «легче», «тяжелее».

- В p разрядах r-ичной системы можно записать r^p чисел. При этом каждая цифра может потребоваться в p экземплярах (скажем, три девятки в 999). Всего заготовленных цифр — вырезанных, например, из картона, — надо иметь $N = p \cdot r$. С помощью этих заготовок можно «записать» $r^{N/r}$ чисел. Функция $r^{N/r}$ достигает максимума при $r = e \approx 2{,}7$. Среди целых чисел максимум обеспечивает $r = 3$. Поэтому иногда говорят, что троичная система счисления — самая экономичная. Двоичная — ей несколько уступает.

[36] При желании обязательно добиться строго максимального результата.

[37] Схема заимствована из статьи Г. Шестопала «Как обнаружить фальшивую монету» (Квант. 1970. **10**).

- **Парадокс Гиббса.** Термодинамика для энтропии идеального газа дает следующую формулу:

$$S = c(T)N \ln V + Ns_0, \tag{8.17}$$

где N — число молекул, V — объем, $c(T)$ — коэффициент, зависящий от температуры T.
Из (8.17) следует, что при смешении двух газов (находящихся в различных объемах V_1 и V_2 при одинаковой температуре) суммарная энтропия возрастает на величину, пропорциональную

$$N_1 \ln \frac{V_1 + V_2}{V_1} + N_2 \ln \frac{V_1 + V_2}{V_2} > 0. \tag{8.18}$$

Для различных газов неравенство (8.18) подтверждается опытом.
Но вывод (8.18) никак не опирается на предположения о сортах смешиваемых газов. Поэтому при смешении одинаковых газов было бы также естественно ожидать возрастания суммарной энтропии. Но для термодинамики это катастрофа, потому что тогда энтропия становится функцией истории газа[38], а не его термодинамического состояния.
Гиббс разрешил противоречие волевым путем, постулировав, что из (8.17) надо вычесть $\ln N! \sim N \ln N$. Тогда энтропия смешения действительно оказывается положительной только для различных газов, и парадокс снимается. Но за введением добавки $\ln N!$, по существу, стоит необходимость отождествлять состояния, получающиеся перестановками молекул, что интуитивно не вполне естественно. Определенным свидетельством того, что здесь не все так просто, может служить список исследователей парадокса: Эйнштейн, Шрёдингер, Планк, Лоренц, Нернст.

[38] Любое состояние газа можно считать полученным в результате устранения ряда перегородок.

Глава 9

Статистика

Статистика — та же теория вероятностей, но — с другого конца.

Из ста миллионов человек опросили тысячу — 777 избирателей за демократию. Какой результат голосования можно прогнозировать, и с какой надежностью? Если выводы малоубедительны, сколько человек надо (было бы) опросить, чтобы прогноз был точным? Или — как контролировать качество продукции, проверяя небольшую часть изделий? Это естественный для статистики круг вопросов.

Отвлекаясь от содержательных интерпретаций, можно сказать так. Статистика — это анализ результатов опыта и определения по ним вероятностных характеристик случайных величин. Такие задачи, безусловно, — в духе **ТВ**. Поэтому статистика для завоевания суверенитета часто настаивает на малосущественных тонкостях и множит количество плохо мотивированных задач, пытаясь создать «армию и пограничные войска». В результате **ТВ**, действующая с меньшей натугой, выглядит привлекательнее. Тем не менее группировка задач вокруг анализа данных опыта заслуживает выделения в самостоятельный раздел. А если при этом соблюдать меру, статистика превращается в симпатичную и полезную ветвь теории вероятностей.

В главе рассматривается идеологическая база статистики в варианте, близком к тезисному. С тяжеловесной частью можно ознакомиться по любому стандартному курсу (см. например, [9, 20]).

9.1. Оценки и характеристики

Основной изучаемой моделью статистики служит многократная реализация случайной величины X. При этом набор независимых случайных величин $X_1, \ldots, X_n$, каждая из которых распределена так же, как и X, — называют *случайной выборкой* [1] объема n. Любую функцию $\Theta_n = \Theta_n(X_1, \ldots, X_n)$ называют *статистической характеристикой* (с. х.), или *статистикой*. Определению обычно подлежат вероятности тех или иных событий, матожидания, дисперсии, корреляции и другие характеристики с. в. на базе с. х.

[1] Иногда выборкой называют реализацию $X_1, \ldots, X_n$.

Например, оценку матожидания m_X можно получить по реализации случайной величины $\Theta_n = (X_1 + \ldots + X_n)/n$, которая в данном случае является одной из возможных с. х. для определения m_X.

Если говорить точнее, то статистика как наука каждый раз вводит гипотезу о вероятностной природе наблюдаемых процессов. Бросается ли монета или берется, скажем, 100 знаков в двоичном разложении числа π, — теория предполагает, что это есть 100-кратная реализация с. в. X, принимающей значения *ноль/один*. Или, скажем, доля леворуких людей равна p. **ТВ** подменяет реальность совсем другой моделью, считая для каждого человека вероятность быть леворуким равной p. Эргодичность (среднее по вероятности равно среднему по реализации) как раз служит основанием адекватности такой модели.

Первое впечатление, что с. х. тривиальны до скуки, отчасти справедливо, — но они далеко не всегда сводятся к примитивному усреднению, как в случае m_X (см. далее). Конечно, статистической характеристикой можно объявить любую функцию $\Theta_n(X_1, \ldots, X_n)$, однако вопрос в том, насколько она удовлетворительна.

Если, например, речь идет об оценке неизвестного параметра θ, характеризующего с. в. X, то оценка θ на основе Θ_n называется *состоятельной*, если $\Theta_n \xrightarrow{P} \theta$ при $n \to \infty$. Из закона больших чисел вытекает состоятельность среднеарифметической оценки матожидания.

В оценках есть также другой существенный аспект. Оценка θ на основе Θ_n называется *смещенной/несмещенной*, если матожидание $\mathbf{E}\{\Theta_n\}$ при любом n *равно/не равно* θ.

Состоятельная оценка не обязана быть несмещенной. (**?**)

Доверительные интервалы. Промежуток, которому принадлежит оцениваемый параметр θ с вероятностью $\geqslant \delta$, называют *доверительным интервалом*, δ — *коэффициентом доверия*, а $1 - \delta$ — *уровнем значимости*.

О справедливости условия

$$\mathbf{P}\{|\theta - \Theta_n| < \varepsilon\} \geqslant \delta,$$

означающего $\theta \in (\Theta_n - \varepsilon, \Theta_n + \varepsilon)$ с вероятностью $\geqslant \delta$, можно судить с помощью неравенства Чебышева, но это даст, конечно,

только грубую оценку. Соответствующий рецепт очевиден. Если θ матожидание X, а Θ_n его несмещенная оценка, то

$$\mathbf{P}\{|\theta - \Theta_n| < \varepsilon\} \geqslant 1 - \frac{\mathbf{D}(\Theta_n)}{\varepsilon^2}.$$

Практический способ действий на этой основе заключается в следующем. Задается коэффициент доверия $\delta = 1 - \mathbf{D}(\Theta_n)/\varepsilon^2$, откуда $\varepsilon = \sqrt{\mathbf{D}(\Theta_n)/(1-\delta)}$, что определяет доверительный интервал $(\Theta_n - \varepsilon, \Theta_n + \varepsilon)$.

В некотором роде здесь заложено противоречие, поскольку на практике обычно имеется реализация выборки и более — ничего. Поэтому в получаемых неравенствах «неизвестное» оценивается через «неизвестное». Дисперсию $\mathbf{D}(\Theta_n)$ приходится определять по той же самой выборке. Однако противоречие снимается, если оценки состоятельны. Тогда $\mathbf{D}(\Theta_n)$ определяется с небольшой ошибкой Δ, и

$$\varepsilon = \sqrt{\frac{1-\delta}{\mathbf{D}(\Theta_n)}} + O(\Delta),$$

т. е. влиянием ошибки при определении дисперсии можно пренебречь.

Если речь идет о достаточно длинных выборках, то можно опираться на предельные теоремы о нормальности распределения ошибок при усреднении, что дает более точные оценки.

Пусть, например, оценивается вероятность p некоторого события A по выборке $X_1, \ldots, X_n$, где X_k принимает значения *ноль/один* в k-м опыте, $X_k = 1$ отвечает «успеху», т. е. наступлению A. Если для оценки используется среднее

$$p_n = \frac{X_1 + \ldots + X_n}{n},$$

то, очевидно,

$$\mathbf{E}\{p_n\} = p, \qquad \mathbf{D}\{p_n\} = \frac{p(1-p)}{n}.$$

При больших n, в силу предельных теорем,

$$\mathbf{P}\left\{|p - p_n| < \varepsilon\sqrt{\frac{p_n(1-p_n)}{n}}\right\} = 2\Phi(\varepsilon), \tag{9.1}$$

откуда получается необходимая связь между крайними точками доверительного интервала и уровнем значимости.

Понятно, что строгое решение (9.1) занимает много места, и приходится кстати, если для диссертации не хватает материала. На самом деле доверительный интервал приближенно равен $(p_n - \varepsilon\sigma, p_n + \varepsilon\sigma)$, где $\sigma = \sqrt{\mathbf{D}\{p_n\}}$.

Но все это хорошо работает, когда выборка достаточно велика (практически, $n \sim 10^2$). При малых n приходится «танцевать» от биномиального распределения, что в отсутствие возможности воспользоваться формулой Стирлинга приводит к весьма громоздким построениям, подталкивающим к графическим методам решения [20].

Оценки матожидания и дисперсии. В случае существования у с. в. X первых двух моментов выборочное среднее

$$\widehat{X}_n = \frac{X_1 + \ldots + X_n}{n},$$

в силу $\mathbf{E}\{\widehat{X}_n\} = m_x$, является несмещенной оценкой. Плюс к тому,

$$\mathbf{D}\{\widehat{X}_n\} = \frac{D_x}{n},$$

что обеспечивает $\widehat{X}_n \xrightarrow{\text{с. к.}} X$, и тем более, $\widehat{X}_n \xrightarrow{P} X$.

Возникает впечатление, что оценка дисперсии

$$\widehat{D}_n = \frac{(X_1 - \widehat{X}_n)^2 + \ldots + (X_n - \widehat{X}_n)^2}{n} \tag{9.2}$$

обладает теми же свойствами, но это не так. Очевидно, после раскрытия в (9.2) получается

$$\widehat{D}_n = \frac{X_1^2 + \ldots + X_n^2 - n\widehat{X}_n^2}{n},$$

откуда

$$\mathbf{E}\{\widehat{D}_n\} = \frac{1}{n} n D_x - \frac{D_x}{n} = \frac{n-1}{n} D_x,$$

что свидетельствует *о смещенности оценки* (9.2). *Несмещенная оценка*:

$$\widehat{D}'_n = \frac{(X_1 - \widehat{X}_n)^2 + \ldots + (X_n - \widehat{X}_n)^2}{n-1}. \tag{9.3}$$

Аккуратный подсчет [20] показывает:

$$\mathbf{D}\{\widehat{D}_n\} = \frac{\mu_4 - \mu_2^2}{n} + \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}, \tag{9.4}$$

откуда ясно, что при существовании центрального четвертого момента μ_4 обе оценки (9.2) и (9.3) состоятельны: $\widehat{D}_n \xrightarrow{P} D_x$, рáвно как и $\widehat{D}'_n \xrightarrow{P} D_x$.

Вопрос о том, какая из оценок (9.2), (9.3) лучше, — однозначного ответа не имеет. Несмещенная оценка точна по матожиданию, но хуже по дисперсии ошибки.

Случайные векторы. В задачах со случайными векторами выборки рассматриваются покоординатно. Новое обстоятельство заключается в появлении смешанных моментов. Но рецептурно все остается по-прежнему.

Например, оценка ковариации

$$\widehat{K}_{xy} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \widehat{X}_n)(Y_k - \widehat{Y}_n)$$

случайного вектора $Z = \{X, Y\}$ — в естественных предположениях состоятельна, но смещена. Несмещенную оценку дает замена в знаменателе n на $n-1$, как и в случае дисперсии.

9.2. Теория и практика

При необходимости проведения, скажем, *опроса населения* — чистый математик оказывается неподготовленным к решению задачи, поскольку на практике существенную роль играют «невероятностные» обстоятельства.

Идет ли речь об опросе избирателей, о социологическом анкетировании или о медицинском обследовании, — из *генеральной совокупности*[2)] необходимо выбрать некоторую долю элементов. Как это сделать? Простейший, казалось бы, вопрос, но на пути его решения очень много препятствий.

Теоретически ситуация выглядит элементарно. Берется полный список, скажем, людей, — и из него равновероятно выбирается какая-то часть населения. Конечно, сама организация случайного выбора — непростая штука, но основные трудности — в другом. Даже общий список с адресами и телефонами может быть проблемой. Список надо достать, завести в память компьютера миллион адресов, обработать.

Проблемы на этом не заканчиваются. После получения в результате случайного отбора списка фамилий приходится «бегать» за каждым респондентом и добиваться от него согласия ответить на вопросы. География случайного выбора оказывается крайне неудачной. В результате — повышенные временные и материальные затраты, *проблема неответивших* и т. п.

Поэтому на практике предпочтение в большинстве случаев отдается более изобретательным технологиям. Можно упомянуть,

[2)] Генеральной совокупностью называют множество всех рассматриваемых элементов. Население города, например.

например, *стратифицированную выборку* с предварительным разбиением генеральной совокупности на группы (*страты*) по какому-либо признаку и последующим случайным отбором внутри групп. Определенный интерес представляют *гнездовые технологии*, в которых случайно выбирается несколько групп с поголовным опросом внутри каждой. Но все это, дрейфуя в сторону эвристики, выходит за рамки статистики как математической науки [3)].

Разумеется, практическая статистика сильно себя скомпрометировала в экономике и социологии. Но даже в этих «скользких» областях она остается единственным средством решения определенного круга задач. В то же время, на фоне иногда анекдотических реалий, математические изыскания «о-малых» выглядят схоластическими. Нельзя, однако, забывать, что есть задачи, где статистика играет совсем другую роль. Оценка физических констант и параметров (на основе многократных измерений), статистическая оптимизация моделей технологических процессов и кое-что еще, где измерения объективны и есть понимание изучаемых процессов.

При этом извлечение максимума возможного во многих ситуациях оказывается принципиальным. Ошибка статистической оценки доли поглощаемых нейтронов приводит к атомному взрыву, а плохая обработка химических анализов поверхностных проб почвы влечет за собой холостое бурение километровых скважин.

9.3. Большие отклонения

Пусть

$$Y_n = \max\{X_1, \ldots, X_n\},$$

где $X_1, \ldots, X_n$ — независимые, одинаково распределенные случайные величины.

Функция распределения Y_n легко определяется,

$$\mathbf{P}\{Y_n \leqslant x\} = \mathbf{P}\bigcap_{i=1}^{n}\{X_i \leqslant x\} = \prod_{i=1}^{n}\mathbf{P}\{X_i \leqslant x\} = F^n(x),$$

[3)] Признать статистику математикой можно, конечно, лишь с той же долей натяжки — что и физику. Статистика начинается там, где теория вероятностей, независимая от устройства Вселенной, начинает увязываться с практикой.

где $F^n(x)$ — общая для всех X_i функция распределения. Понятно, что распределение с. в. Y_n определяет поведение «правого хвоста» $F(x)$.

Например, при условии

$$\lim_{x\to\infty}[1 - F(x)]x^a = b, \quad a, b > 0,$$

«нормированная» с. в. $\zeta_n = Y_n/(bn)^{1/a}$ при $n \to \infty$ сходится по распределению к с. в. ζ,

$$\mathbf{P}\{\zeta \leqslant x\} = \begin{cases} e^{-x^{-a}}, & x > 0; \\ 0, & x \leqslant 0. \end{cases} \quad (?)$$

Статистика успехов в схеме Бернулли. При изучении суммы

$$S_k = X_1 + \ldots + X_k,$$

где «успех» $X_n = 1$ достигается с вероятностью p, соответственно $X_n = 0$ — с вероятностью $1 - p$, — удобно рассматривать *нормированную сумму*

$$\widehat{S}_k = \frac{S_k - kp}{\sigma\sqrt{k}}, \qquad \sigma^2 = p(1-p).$$

Теорема Муавра—Лапласа гарантирует сходимость $\widehat{S}_k$ к нормальному распределению $\mathcal{N}(0, 1)$, что означает

$$\lim_{n\to\infty} \mathbf{P}\{\alpha \leqslant \widehat{S}_k \leqslant \beta\} = \Phi(\beta) - \Phi(\alpha). \tag{9.5}$$

На практике, естественно, возникает вопрос о соотношении k и границ доверительного интервала, при котором из (9.5) $\lim\limits_{k\to\infty}$ можно убрать, не слишком нарушая равенство.

Критерием здесь может служить величина $\dfrac{x^3}{\sigma\sqrt{k}} < \varepsilon$. При больших x и k, но малых ε,

$$\mathbf{P}\{x \leqslant \widehat{S}_k\} \approx 1 - \Phi(x) \approx \frac{e^{-x^2/2}}{x\sqrt{2\pi}},$$

что подтверждается несложными выкладками [26].

При малом объеме выборки приходится пользоваться для оценки S_k точным биномиальным распределением, но тогда возникают неудобства счета, снова ведущие к огрубленным оценкам.

Закон повторного логарифма. Идеологически другая задача возникает при попытке оценить поведение возможных траекторий $\widehat{S}_k$ в целом. Несмотря на нулевое матожидание и единичную дисперсию с. в. $\widehat{S}_k$ (при любом k), — в любой типичной реализации последовательности $\widehat{S}_1, \widehat{S}_2, \ldots$ будут встречаться сколь угодно большие значения. Из этого расплывчатого соображения можно извлечь точную закономерность: *верхний предел*

$$\overline{\lim_{k\to\infty}} \frac{\widehat{S}_k}{\sqrt{\ln\ln k}} = \sqrt{2} \tag{9.6}$$

с вероятностью единица.

Это так называемый *закон повторного логарифма Хинчина.*

Если вернуться непосредственно к сумме S_k, то (9.6) означает, что при достаточно больших k типичные траектории не выходят за пределы $S_k \leqslant kp + \sigma\sqrt{2k\ln\ln k}$.

9.4. От «хи-квадрат» до Стьюдента

Хи-квадрат распределение имеет плотность $\rho(x) = 0$ при $x \leqslant 0$ и

$$\rho(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} \quad \text{при} \quad x > 0,$$

где Γ — гамма-функция[4], а целочисленный параметр n называют *числом степеней свободы.*

Так распределен квадрат вектора $\chi = \{X_1, \ldots, X_n\}$,

$$\chi^2 = X_1^2 + \ldots + X_n^2,$$

с нормальными координатами X_k, имеющими нулевые матожидания и единичные дисперсии.

При $n = 2$ χ^2*-распределение* совпадает с показательным.

Распределение Стьюдента (t-распределение) имеет случайная величина

$$t = \frac{\sqrt{n}X}{\sqrt{\chi^2}},$$

[4] $\Gamma(\nu) = \int\limits_0^\infty x^{\nu-1} e^{-x}\, dx.$

где n — *число степеней свободы*, с. в. X имеет нормальное распределение $\mathcal{N}(0, 1)$, а χ^2 распределена по закону «хи-квадрат».

Распределение t,

$$\mathbf{P}\{t < x\} = \frac{\Gamma((n+1)/2)}{\sqrt{2\pi}\,\Gamma(n/2)} \int_{-\infty}^{x} \left(1 + \frac{s^2}{n}\right)^{-(n+1)/2} ds,$$

не очень подходит для запоминания. Но по таблицам — при необходимости — считается, а при больших n мало отличается от $\mathcal{N}(0, 1)$, и слабо сходится к $\mathcal{N}(0, 1)$ при $n \to \infty$.

На практике распределение Стьюдента широко применяется[5] в следующей стандартной схеме. Для независимых $X_1, \ldots, X_n$, распределенных по закону $\mathcal{N}(m, \sigma^2)$, лучшие несмещенные оценки m и σ^2 дают статистики

$$\widehat{X}_n = \frac{X_1 + \ldots + X_n}{n} \quad \text{и} \quad \widehat{D}_n = \frac{(X_1 - \widehat{X}_n)^2 + \ldots + (X_n - \widehat{X}_n)^2}{n-1}.$$

При этом $(\widehat{X}_n - m)/\sigma$ подчиняется закону $\mathcal{N}(0, 1)$, с. в. $(n-1)\widehat{D}_n/\sigma^2 = \chi^2$ — закону «хи-квадрат», а отношение

$$\frac{(\widehat{X}_n - m)/\sigma}{\sqrt{\widehat{D}_n/\sigma^2}} = \sqrt{n-1}\,\frac{\widehat{X}_n - m}{\sqrt{\chi^2}}$$

оказывается распределенным по Стьюденту, что позволяет более точно[6] оценивать доверительные интервалы по заданному уровню значимости (см. [20, 21]). Некоторые тонкости, оставшиеся здесь за кадром, требуют громоздких выкладок при незначительном «идеологическом эффекте».

9.5. Максимальное правдоподобие

Главная беда статистики в расхождении слова и дела. Громоздкие формулы сопровождаются приговариванием о необходимости высокой точности, тогда как всем ясно, что неучтенные факторы перекрывают любые математические усилия. Поэтому оправдывать изыскания более естественно красотой самих задач.

Случайные величины $X_1, \ldots, X_n$ независимы и одинаково распределены с плотностью $\rho_\theta(x)$. Необходимо изобрести наилучшую оценку

$$\Theta_n = \Theta_n(X_1, \ldots, X_n)$$

параметра θ.

---

[5] По крайней мере, так говорят.

[6] По сравнению с грубыми оценками, базирующимися на неравенстве Чебышева.

Вполне нормальная математическая проблема, способная инициировать поток диссертаций и разговоры о практической значимости, необходимые для защиты территории. Червоточинка заключена в большой неоднозначночности понятия «наилучшая», не говоря о том, что разногласия начинаются уже на этапе «приемлемости». О противоречивости несмещенности и состоятельности уже говорилось. Но на эту мельницу воду можно лить и лить.

Метод максимального правдоподобия Фишера заключается в максимизации совместной плотности[7]

$$\rho_\theta(x_1, \ldots, x_n) = \prod_{k=1}^{n} \rho_\theta(x_k)$$

распределения $X_1, \ldots, X_n$ при полученных реализациях $x_1, \ldots, x_n$. Функция $\Theta_n = \Theta_n(x_1, \ldots, x_n)$, обеспечивающая такой максимум, называется *оценкой максимального правдоподобия*.

Идея вполне изящная, и в простых ситуациях[8] хорошо работает. Но кругом «мины», о которых легко догадаться, поскольку распределения и установки оптимизации могут быть другими. Естественно, например, минимизировать дисперсию $\mathbf{E}\{(\Theta_n - \theta)^2\}$, что в общем случае будет приводить к иным решениям. Обнаруживается, правда, интересный факт — *неравенство Рао—Крамера*:

$$\mathbf{E}\{(\Theta_n - \theta)^2\} \geqslant \frac{1}{I_n(\theta)}, \tag{9.7}$$

где

$$I_n(\theta) = \mathbf{E}\left\{\frac{\partial}{\partial\theta} \ln \rho_\theta(X_1, \ldots, X_n)\right\}$$

— *количество информации по Фишеру*.

При обращении (9.7) в равенство оценку называют *эффективной*. Но это бывает, как говорят, после дождичка в четверг — что чаще всего соответствует одномерному нормальному распределению.

Достаточные статистики — другая идея Фишера. В случае, например, нормального распределения $\rho_\theta(x)$ со средним θ вся информация о θ содержится в оценке

$$\Theta_n = \frac{X_1 + \ldots + X_n}{n},$$

что вытекает из независимости распределения

$$\{X_1 - \Theta_n, \ldots, X_n - \Theta_n\}.$$

---

[7] Которая в случае зависимых $X_1, \ldots, X_n$ не обязательно равна произведению $\rho_\theta(x_k)$.

[8] Типа оценки вероятности в схеме Бернулли либо параметра распределения Пуассона.

Это означает, что после подсчета среднего арифметического выборки — сама выборка перестает быть нужной, из нее уже ничего дополнительно не выжмешь. При этом ясно, что при наличии обратимой функциональной связи $\Upsilon_n = \varphi(\Theta_n)$ статистика Υ_n остается достаточной, поскольку Θ_n можно восстановить. Это замечание подчеркивает тот факт, что в «достаточности» определяющую роль играет присутствие в оценке *всей информации*. Как конкретно оценивать — другой вопрос.

В общем случае набор функций

$$S_1(X_1, \ldots, X_n), \ldots, S_k(X_1, \ldots, X_n)$$

считается достаточной статистикой относительно θ, если совместное распределение $X_1, \ldots, X_n$ при фиксированных $S_1, \ldots, S_k$ не зависит от θ.

Определять, конечно, легко — пользоваться трудно.

Парадокс Фишера. Для двумерного нормального распределения с независимыми координатами, имеющими единичные дисперсии и неизвестные матожидания θ_1, θ_2, — обычное среднее $\{\Theta_1, \Theta_2\}$ двумерной выборки является достаточной статистикой для пары $\{\theta_1, \theta_2\}$.

Вектор $\{\theta_1, \theta_2\}$ можно описывать в полярных координатах (r, θ), оценивая θ по тангенсу $\operatorname{tg}(\Theta_2/\Theta_1)$, а r по величине $\sqrt{\Theta_1^2 + \Theta_2^2}$.

В силу взаимной однозначности декартовых и полярных координат оба варианта статистики достаточны. Но распределение $\sqrt{\Theta_1^2 + \Theta_2^2}$ из-за сферической симметрии относительно точки $\{\theta_1, \theta_2\}$ не зависит от θ. Отсюда (вроде бы) следует, что информация об r ничего не добавляет к информации о θ. В то же время ясно, что $\mathbf{E}\{(\Theta - \theta)^2 | r\}$ зависит от r (тем сильнее, чем меньше r), и это легко подтверждается вычислением.

Главная причина видимого противоречия заключена в определении достаточности. Информация не потеряна, но отсюда вовсе не следует, что

$$\Theta = \operatorname{arctg} \frac{\Theta_2}{\Theta_1}$$

— хорошая оценка. Почему бы, например, оценивая θ, не танцевать от синуса отношения $\Theta_2/\sqrt{\Theta_1^2 + \Theta_2^2}$?

9.6. Парадоксы

Парадоксы — ценная вещь в том смысле, что без них нет границ. Нож теории идет как в масло, и создается впечатление, что Вселенная — изотропно-масляная во всех направлениях.

Оазисы противоречий разнообразят движение и способствуют прозрению. В статистике, правда, слишком много парадоксов небольшого калибра, да еще основанных на искусственных понятиях.

Парадокс Стейна[9]. Оценка матожиданий трехмерного нормального закона по обычному среднему трехмерной выборки — *не допустима.*

Звучит торжественно. Примерно как нарушение закона сохранения энергии. И даже жалко разъяснять, что недопустимые статистики вполне приемлемы. *Допустимая статистика* Θ^* происходит из категории минимаксных оценок и минимизирует «потери» при наихудшем θ, т. е.

$$\sup_{\theta} \mathbf{E}\{(\Theta^* - \theta)^2\} = \inf_{\Theta} \sup_{\theta} \mathbf{E}\{(\Theta - \theta)^2\}.$$

Конечно, это имеет смысл, но беда статистики в том, что в ней имеют смысл миллион понятий, которые не очень хорошо согласуются друг с другом. Многое из придуманного идеально подходит только для нормально распределенных величин[10]. А тут Стейн — с неприятным сюрпризом. Оказывается, в «нормальном» случае тоже не все нормально. Штуковина, безусловно, калибра «о-малого», но как идеологическая диверсия — весома. «Разбор полетов» см. у Секея [22].

С практической точки зрения бóльшую ценность имеют противоречия, демонстрирующие слабые звенья интуиции. Скажем, ловушку из раздела 1.2 вполне можно включать в школьную программу. Ничего сложного, но обман приходится как раз на ахиллесову пяту здравого смысла, — что обеспечивает лечебный эффект. Главное ведь не в толщине маскирующего слоя из формул, а в точности попадания в незащищенные места подсознания. Поэтому, если на парадоксы смотреть как на таблетки от хронической беспечности, то заботиться надо о соответствии диагнозу.

Распространенная причина заблуждений — перенос старого опыта на новые понятия. В ядре многих статистических казусов лежат противоестественные свойства вероятностных неравенств.

Сравнивая случайные величины X и Y, пишут $X < Y$ и говорят *«X меньше Y по вероятности»*, подразумевая

$$\mathbf{P}\{X < Y\} > \mathbf{P}\{X \geqslant Y\}, \tag{9.8}$$

т. е. вероятность неравенства события $\{X < Y\}$ больше $1/2$.

---

[9] *Stein C.* // Proc. Third Berkeley Symp. on Math. Stat. & Prob. 1956. **1**. P. 197–206.

[10] Даже среднее для оценки матожидания в «не нормальном» случае — не такая уж хорошая оценка.

Это совсем непохоже на $7 > 3$, но возникающие ассоциации путают карты, порождая неприятности типа *парадокса транзитивности* (раздел 1.5). Свойства отношения (9.8) настолько сильно отличаются от обычного неравенства, что знак $<$ имело бы смысл заменить каким-либо иным.

Неразбериху усугубляют другие понятия *больше/меньше* для с. в. Говорят, например, «X *меньше* Y *стохастически*», если

$$\mathbf{P}\{X < \tau\} \geqslant \mathbf{P}\{Y < \tau\}, \quad \text{причем} \quad \mathbf{P}\{X < \tau_0\} > \mathbf{P}\{Y < \tau_0\}$$

для некоторого τ_0.

Оба понятия, вроде бы, естественным образом мотивированы, но могут сильно отличаться друг от друга. Например, если Y равномерно распределена на $[0, 1]$, а

$$X = \begin{cases} \varepsilon^2 Y, & \text{с вероятностью } \varepsilon, \\ Y + \varepsilon^2(1 - Y), & \text{с вероятностью } 1 - \varepsilon, \end{cases}$$

то при малых $\varepsilon > 0$

$$X > Y \quad \textit{с вероятностью } 1 - \varepsilon, \textit{ но } \quad X < Y \quad \textit{стохастически.}$$

К сказанному можно многое добавить, но это больше подходит для самостоятельной тренировки, потому что дело не в отдельных парадоксах. Вероятностные неравенства — это главный механизм всех неприятностей в **ТВ**. Главный и постоянно действующий, ибо события как подмножества Ω описываются неравенствами, — и ни одна задача не обходится без анализа «равно-больше-меньше». Поэтому без привычки к двойственному характеру неравенств, ограничивающих множества и сравнивающих меры, ориентироваться в **ТВ** трудно. И что хуже всего, требуется умение держать внимание на двух факторах одновременно, для чего надо быть Юлием Цезарем.

Неравенства в задачах **ТВ** иногда лежат на поверхности, но чаще — за кадром. Вот классический пример.

Парадокс Эджворта (XIX в.) — изюминка из разряда «чем больше данных, тем хуже результат». Практического значения, можно сказать, не имеет, но факт — принципиальный.

Речь идет о возможности неравенства

$$\rho_X(0) > \rho_{\widehat{X}}(0). \tag{9.9}$$

Здесь ρ_X плотность с. в. X, а $\rho_{\widehat{X}}$ плотность $\widehat{X} = (X_1 + X_2)/2$, где X_1 и X_2 независимы и одинаково распределены с X.

Неприятность (9.9) обеспечивает плотность

$$\rho_x(x) = \frac{3}{2(1+|x|)^4}.$$

В результате $\mathbf{P}\{|X| < \varepsilon\} > \mathbf{P}\{|\widehat{X}| < \varepsilon\}$ при малых[11] $\varepsilon > 0$, что как раз означает ухудшение оценки нулевого матожидания X при увеличении объема выборки с 1 до 2.

[11] На самом деле годится любое $\varepsilon > 0$.

Глава 10

Сводка основных определений и результатов

10.1. Основные понятия

✓ *Вероятностным пространством* называется непустое множество Ω с «узаконенным» семейством $\mathcal{A}$ его подмножеств и неотрицательной функцией (мерой) $\mathbf{P}$, определенной на $\mathcal{A}$ и удовлетворяющей условию $\mathbf{P}(\Omega) = 1$, а также

$$\mathbf{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

для любой последовательности $A_1, A_2, \ldots \in \mathcal{A}$ взаимно непересекающихся множеств A_i. Другими словами, вероятностное пространство определяет тройка $(\Omega, \mathcal{A}, \mathbf{P})$.

✓ На роль $\mathcal{A}$ годится не любое семейство, поскольку сумма и пересечение событий не должны выводить за рамки дозволенного. Но тогда приходится требовать

$$A, B \subset \mathcal{A} \quad \Rightarrow \quad A \cup B \subset \mathcal{A}, \quad A \cap B \subset \mathcal{A},$$

дополнительно оговаривая $\Omega \in \mathcal{A}$, и принадлежность $\mathcal{A}$ любого A вместе с дополнением. Такая совокупность множеств называется *алгеброй* подмножеств Ω, и — *σ-алгеброй*, в более общем случае, когда в $\mathcal{A}$ входят любые суммы и пересечения *счетных совокупностей* $A_k \subset \mathcal{A}$:

$$\bigcup_{k}^{\infty} A_k \subset \mathcal{A}, \qquad \bigcap_{k}^{\infty} A_k \subset \mathcal{A}.$$

✓ Множество Ω с заданной на нем σ-алгеброй $\mathcal{A}$ называют *измеримым пространством*. В случае, когда Ω представляет собой вещественную прямую, — *борелевская σ-алгебра* $\mathcal{B}$ порождается [1] системой непересекающихся полуинтервалов $(\alpha, \beta]$. Элементы $\mathcal{B}$ называют *борелевскими множествами*.

В случае $\Omega = R^n$ борелевские множества определяются аналогично (как прямые произведения одномерных).

---

[1] Взятием всевозможных объединений и пересечений.

✓ Вещественная функция $f(\omega)$ называется *измеримой* относительно σ-алгебры $\mathcal{A}$, если прообраз любого борелевского множества принадлежит $\mathcal{A}$. Если $\mathcal{A} = \mathcal{B}$, говорят, что функция $f(\omega)$ — *борелевская*.

✓ Любое подмножество $A \in \mathcal{A}$ называют событием A, а его меру $\mathbf{P}(A)$ — вероятностью события A.

Объединением или *суммой событий* A и B называют событие, состоящее в наступлении хотя бы одного из событий A, B, и обозначаемое как $A \cup B$ или $A + B$. Первое обозначение прямо указывает, какое множество в Ω отвечает сумме событий.

Пересечением или *произведением событий* A и B называют событие, состоящее в совместном наступлении A, B, и обозначаемое как $A \cap B$ или AB.

$$\boxed{\mathbf{P}(A + B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(AB).}$$

✓ Событию «не A» отвечает дополнение $\overline{A}$ множества A в Ω, а разность $A \setminus B$, или $A - B$, интерпретируется как наступление A, но не B. Наконец, *симметрическая разность*

$$A \triangle B = (A \cup B) \setminus (A \cap B)$$

обозначает событие, состоящее в наступлении одного из A, B, но не двух вместе.

Пустое множество $\varnothing$, считается, принадлежит Ω и символизирует невозможное событие. При этом $\mathbf{P}(\varnothing) = 0$.

✓ Вероятность $\mathbf{P}(B|A)$ наступления B при условии наступления A — называют *условной*,

$$\boxed{\mathbf{P}(B|A) = \frac{\mathbf{P}(AB)}{\mathbf{P}(A)},}$$

откуда

$$\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B|A),$$

что именуют *формулой умножения вероятностей*.

✓ Разбиение Ω на *полную группу несовместимых* (*непересекающихся*) *событий* $A_1, \ldots, A_n$ позволяет любое событие B записать в виде

$$B = BA_1 + \ldots + BA_n,$$

откуда $\mathbf{P}(B) = \mathbf{P}(BA_1) + \ldots + \mathbf{P}(BA_n)$, что приводит к *формуле полной вероятности*:

$$\boxed{\mathbf{P}(B) = \mathbf{P}(B|A_1)\mathbf{P}(A_1) + \ldots + \mathbf{P}(B|A_n)\mathbf{P}(A_n).}$$

Формула Байеса,

$$\mathbf{P}(A_j|B) = \frac{\mathbf{P}(B|A_j)\mathbf{P}(A_j)}{\sum\limits_k \mathbf{P}(B|A_k)\mathbf{P}(A_k)},$$

интерпретируется как правило определения апостериорных вероятностей $\mathbf{P}(A_j|B)$ по априорным $\mathbf{P}(A_k)$.

✓ *Случайной величиной* называется любая вещественная $\mathcal{A}$-измеримая функция $X(\omega)$, заданная на $(\Omega, \mathcal{A})$.

✓ *Среднее значение* $m_x = \mathbf{E}(X)$,

$$\boxed{\mathbf{E}(X) = \sum_{\omega\in\Omega} X(\omega)\mathbf{P}(\omega)}$$

называют *матожиданием* $X(\omega)$.

Математическое ожидание *функции-индикатора* $\chi_A(\omega)$ множества A,

$$\chi_A(\omega) = \begin{cases} 1, & \text{если } \omega \in A; \\ 0, & \text{если } \omega \notin A, \end{cases}$$

равно вероятности $\mathbf{P}(A)$.

$$\mathbf{E}(\alpha X + \beta Y) = \alpha\mathbf{E}(X) + \beta\mathbf{E}(Y).$$

✓ В континуальном случае, когда точки в Ω распределены с плотностью $\mu(\omega)$, причем $\int\limits_\Omega \mu(\omega)\,d\omega = 1$, — вероятность события $\omega \in A$ определяется как

$$\mathbf{P}(A) = \int\limits_A \mu(\omega)\,d\omega,$$

а если на Ω задана случайная величина $X(\omega)$, в том числе векторная $X(\omega) = \{X_1(\omega), \ldots, X_n(\omega)\}$, то матожидание равно

$$\mathbf{E}(X) = \int\limits_\Omega X(\omega)\mu(\omega)\,d\omega.$$

✓ Случайные величины X часто описывают с помощью *функции распределения*:

$$\boxed{F(x) = P(X < x).}$$

При этом отказ от рассмотрения исходного пространства элементарных событий носит условный характер. Просто одно пространство заменяется другим. Новым пространством Ω случайной величины X становится вещественная прямая или ее подмножество.

Функция $F(x)$ монотонно возрастает (не убывает) и

$$\lim_{x\to\infty} F(x) = 1, \qquad \lim_{x\to-\infty} F(x) = 0.$$

✓ Наряду с $F(x)$ используется также *плотность распределения* $\rho(x)$, связанная с $F(x)$ условием:

$$F(x) = \int_{-\infty}^{x} \rho(u)\, du,$$

откуда

$$\rho(x) = F'(x).$$

✓ **Независимость.** События A и B называют *независимыми*, если $\mathbf{P}(B|A) = \mathbf{P}(B)$ (равносильно $\mathbf{P}(A|B) = \mathbf{P}(A)$), т. е. формула умножения вероятностей переходит в $\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B)$.

Определение n независимых событий имеет вид

$$\mathbf{P}(A_1 \dots A_n) = \mathbf{P}(A_1) \dots \mathbf{P}(A_n).$$

В применении к случайному вектору с независимыми компонентами $X = \{X_1, X_2\}$ это дает

$$\mathbf{P}(X_1 < x_1, X_2 < x_2) = \mathbf{P}(X_1 < x_1)\mathbf{P}(X_2 < x_2),$$

что влечет за собой

$$\boxed{F(x_1, x_2) = F_1(x_1)F_2(x_2),}$$

и, как следствие,

$$\boxed{\rho(x_1, x_2) = \rho_1(x_1)\rho_2(x_2).}$$

✓ Скаляр

$$\boxed{\mathbf{D}(X) = \mathbf{E}(X - m_x)^2}$$

называется *дисперсией* случайной величины X, а $\sigma_x = \sqrt{\mathbf{D}(X)}$ — *среднеквадратическим отклонением* X от своего среднего значения m_x.

✓ Для двух случайных величин X, Y рассматривают *смешанные моменты* $\mathbf{E}(X^n Y^m)$. Важную роль во многих ситуациях играют *ковариация*

$$\boxed{\operatorname{cov}(XY) = \mathbf{E}[(X - m_x)(Y - m_y)]}$$

и *коэффициент корреляции*

$$r_{xy} = \frac{\operatorname{cov}(XY)}{\sigma_x \sigma_y}.$$

✓ **Неравенство Коши—Буняковского:**

$$\boxed{\mathbf{E}(|XY|) \leqslant \sqrt{\mathbf{E}(X^2)\mathbf{E}(Y^2)}.}$$

✓ **Неравенство Чебышева**:

$$\boxed{\mathrm{P}(|X - m_x| \geqslant a) \leqslant \frac{\mathrm{D}(X)}{a^2}.}$$

✓ **Неравенство Колмогорова.** *Пусть последовательность независимых случайных величин* X_j *имеет нулевые матожидания* $\mathrm{E}\{X_j\} = 0$ *и* $\mathrm{D}\{X_j\} < \infty$. *Тогда*

$$\boxed{\mathrm{P}\Big\{\max_{k \leqslant n} |X_1 + \ldots + X_k| \geqslant \varepsilon\Big\} \leqslant \frac{1}{\varepsilon^2} \sum_{j=1}^{n} \mathrm{D}\{X_j\}.}$$

✓ **Неравенство Иенсена.** *Пусть* $\varphi(x)$ — *вогнутая функция* (*выпуклая вверх*), *и матожидание* $\mathrm{E}(X)$ *существует. Тогда*

$$\boxed{\mathrm{E}\{\varphi(X)\} \leqslant \varphi(\mathrm{E}\{X\}).}$$

10.2. Распределения

✓ **Равномерное распределение** в промежутке $[a, b]$ имеет плотность

$$\rho(x) = \frac{1}{b - a},$$

которой соответствует функция распределения

$$F(x) = \int_{-\infty}^{x} \rho(u)\, du = \frac{1}{b - a} \int_{a}^{x} du = \frac{x - a}{b - a}.$$

✓ *Биномиальное распределение*

$$\boxed{p_k = C_n^k p^k q^{n-k}}$$

имеет сумма

$$S_n = X_1 + \ldots + X_n,$$

где все с. в. X_k независимы и принимают два возможных значения 1 или 0 с вероятностями p и $q = 1 - p$. Сумма принимает значение $S_n = k$ с вероятностью p_k.

Легко проверяется:

$$\mathrm{E}\{S_n\} = np, \qquad \mathrm{D}\{S_n\} = np(1 - p), \qquad \mathrm{E}\{[S_n - np]^3\} = np(1 - p)(1 - 2p).$$

✓ Вероятность появления k нулей перед первым появлением единицы равна $\boxed{p_k = pq^k}$. Совокупность этих вероятностей (при $k = 0, 1, 2, \ldots$) называют

геометрическим распределением. Геометрическое распределение имеет случайная величина, равная числу испытаний до первого успеха.

✓ **Распределение Пуассона**, как и биномиальное, является дискретным, и характеризуется вероятностями

$$\boxed{\mathbf{P}(X = k) = \frac{a^k}{k!}e^{-a}} \qquad (k = 0, 1, \ldots).$$

Вычисление показывает, что

$$a = \sum_{k=0}^{\infty} k\mathbf{P}(X = k),$$

т. е. параметр a есть матожидание с. в. X, распределенной по закону Пуассона. Дисперсия X тоже равна a.

✓ **Нормальный закон распределения**, обозначаемый обычно как $\mathcal{N}(m_x, \sigma_x^2)$ имеет плотность

$$\rho(x) = \frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{(x-m_x)^2}{2\sigma_x^2}},$$

однозначно определяемую матожиданием m_x и дисперсией σ_x^2.

Функция распределения $\mathcal{N}(0, 1)$ имеет вид

$$\boxed{\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-s^2/2}\,ds.}$$

✓ **Функции случайных величин.** Если $Y = f(X)$, где f обычная детерминированная функция, а X случайная величина с плотностью $\rho(x)$, то

$$m_y = \mathbf{E}(Y) = \int f(x)\rho(x)\,dx.$$

Аналогично,

$$\sigma_y^2 = \mathbf{D}(Y) = \int [f(x) - m_y]^2\rho(x)\,dx.$$

Если $Y = f(X)$ вектор, подобным образом определяются и ковариации:

$$\operatorname{cov}(Y_iY_j) = \int [f_i(x) - m_{y_i}][f_j(x) - m_{y_j}]\rho(x)\,dx.$$

✓ Если $f(x)$ обратима, то

$$F(y) = \mathbf{P}\{Y < y\} = \int_{-\infty}^{y} \rho(f^{-1}(y))\,|[f^{-1}(y)]'|\,dy,$$

а плотность

$$\rho_y(y) = \rho_x(f^{-1}(y))\left|[f^{-1}(y)]'\right|,$$

где индексы x, y показывают, какие плотности подразумеваются.

С той же целью может быть использована формула

$$\rho_y(y) = \int \rho_x(x)\delta[y - f(x)]\,dx.$$

Если X и Y — векторы, имеющие одинаковую размерность, и

$$X = f^{-1}(Y) = h(Y),$$

то

$$F(y) = \mathrm{P}\{Y < y\} = \int_{-\infty}^{y_1} \dots \int_{-\infty}^{y_n} \rho_x(h(y)) \det\left[\frac{\partial h_i}{\partial y_j}\right] dy,$$

где

$$\rho_x(h(y)) \det\left[\frac{\partial h_i}{\partial y_j}\right] = \rho_y(y).$$

✓ При известной совместной функции распределения

$$F(u, v) = \mathrm{P}\{U < u, V < v\}$$

случайного вектора $X = \{u, v\}$, имеем

$$F_u(u) = F(u, \infty), \qquad F_v(v) = F(\infty, v).$$

Соответственно,

$$\rho_u(u) = \int_{-\infty}^{\infty} \rho(u, v)\,dv, \qquad \rho_v(v) = \int_{-\infty}^{\infty} \rho(u, v)\,du.$$

✓ Формула для условной плотности распределения:

$$\rho(y|x) = \frac{\rho(x, y)}{\rho(x)},$$

откуда $\rho(x, y) = \rho(y|x)\rho(x)$.

✓ Через условную плотность определяются любые условные моменты, в том числе *условное матожидание*:

$$\mathrm{E}(Y|x) = \int y\rho(y|x)\,dy.$$

✓ Функция $\varphi(\lambda) = \mathbf{E}\,(e^{i\lambda X})$ называется *характеристической функцией* с. в. $\boldsymbol{X}$. При записи

$$\varphi(\lambda) = \int_{-\infty}^{\infty} \rho(x) e^{i\lambda x}\, dx, \qquad i^2 = -1, \qquad -\infty < \lambda < \infty$$

это в несущественных деталях отличается от стандартного *преобразования Фурье* плотности $\rho(x)$.

✓ При условии абсолютной интегрируемости

$$\int |\varphi(\lambda)|\, d\lambda < \infty,$$

соответствующая плотность однозначно восстанавливается «обратным преобразованием Фурье»

$$\rho(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(\lambda) e^{-i\lambda x}\, d\lambda.$$

Если с. в. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ независимы, то х. ф. $\varphi(\lambda)$ суммы $\boldsymbol{X}_1 + \ldots + \boldsymbol{X}_n$ равна произведению х. ф. слагаемых:

$$\varphi(\lambda) = \prod_k \varphi_k(\lambda).$$

Это обстоятельство и определяет заметную роль характеристических функций в теории вероятностей.

✓ Вот характеристические функции стандартных распределений.

распределение	*плотность*	*х. ф.* $\varphi(\lambda)$
нормальное	$\rho(x) = \dfrac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}$	$e^{im_x\lambda - \frac{1}{2}\sigma_x^2\lambda^2}$
равномерное	$\rho(x) = \dfrac{1}{b-a}$ на $[a, b]$	$\dfrac{e^{i\lambda b} - e^{i\lambda a}}{i\lambda(b-a)}$
Коши	$\rho(x) = \dfrac{a}{\pi(x^2 + a^2)}$	$e^{-a\lvert\lambda\rvert}$
показательное	$\rho(x) = ae^{-ax},\ x \geqslant 0$	$\dfrac{a}{a - i\lambda}$
показательное-2	$\rho(x) = \dfrac{1}{2} e^{-\lvert x\rvert}$	$\dfrac{1}{1+\lambda^2}$

✓ Если случайная величина X принимает дискретные значения $X = k$ с вероятностями p_k, то

$$\Pi(z) = \sum_{k=0}^{\infty} p_k z^k$$

называют *производящей функцией* с. в. X. В общем случае целочисленной случайной величины X производящая функция

$$\Pi(z) = \mathrm{E}\{z^X\}.$$

д-распределение	*д-плотность*	*х. ф.* $\varphi(\lambda)$	*п. ф.* $\Pi(z)$
биномиальное	$p_k = C_n^k p^k q^{n-k}$	$(pe^{i\lambda} + q)^n$	$(pz + q)^n$
геометрическое	$p_k = pq^k$	$\dfrac{p}{1 - qe^{i\lambda}}$	$\dfrac{p}{1 - qz}$
Пуассоновское	$p_k = \dfrac{a^k}{k!} e^{-a}$	$e^{a(e^{i\lambda} - 1)}$	$e^{-a(1-z)}$

С характеристической функцией $\varphi(\lambda)$ ее связывает соотношение

$$\varphi(\lambda) = \Pi(e^{i\lambda}).$$

10.3. Законы больших чисел

✓ *Пусть некоррелированные случайные величины X_i имеют одно и то же математическое ожидание μ и одну и ту же дисперсию σ^2. Тогда при любом $\varepsilon > 0$*

$$\mathrm{P}\left\{\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| > \varepsilon\right\} \leqslant \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \textit{при} \quad n \to \infty.$$

Это один из вариантов слабого закона больших чисел, в котором речь идет о стабилизации с. в. $\dfrac{S_n}{n} = \dfrac{X_1 + \ldots + X_n}{n}$.

Предположения о том, что величины X_i имеют одинаковые математические ожидания и дисперсии — необязательны.

✓ **Усиленный закон больших чисел.** Один из вариантов: *Пусть независимые величины X_n имеют математические ожидания μ_n и дисперсии σ_n^2. При условии* $\displaystyle\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty$

$$\frac{X_1 + \ldots + X_n}{n} - \frac{\mu_1 + \ldots + \mu_n}{n} \to 0 \qquad (n \to \infty)$$

с вероятностью единица.

✓ Последовательность функций $f_n(x)$ *асимптотически постоянна*, если существует такая числовая последовательность μ_n, что

$$\mathbf{P}\{|f_n(x) - \mu_n| > \varepsilon\} \to 0 \quad \text{при} \quad n \to \infty$$

для любого наперед заданного $\varepsilon > 0$. Либо, в более жестком варианте,

$$\mathbf{D}\{f_n(x)\} \to 0 \quad \text{при} \quad n \to \infty.$$

✓ **Теорема.** *Пусть независимые с. в. X_i распределены на $[0, 1]$ с плотностями $\rho_i(x_i)$, причем все $\rho_i(x_i) > \varepsilon > 0$, а последовательность функций $f_n(x_1, \ldots, x_n)$ удовлетворяет неравенствам*

$$\|\nabla f_n(x)\| \leqslant \gamma_n, \quad x \in C_n.$$

Тогда при $\gamma_n < \gamma < \infty$ дисперсия $\mathbf{D}\{f_n\}$ ограничена некоторой константой, не зависящей от n. Если же γ_n стремится к нулю с ростом n, то

$$\mathbf{D}\{f_n\} \to 0 \quad \textit{при} \quad n \to \infty,$$

т. е. последовательность функций $f_n(x)$ асимптотически постоянна на C_n.

✓ В общем случае для оценки дисперсий нелинейных функций оказывается эффективен следующий результат, опирающийся на понятие *сопряженной плотности*:

$$\rho^*(x) = \mu(\infty) \int_{-\infty}^{x} \rho(t)\, dt - \mu(x), \qquad \mu(x) = \int_{-\infty}^{x} t\rho(t)\, dt.$$

Лемма. *Пусть независимые с. в. X_i распределены независимо с плотностями $\rho_i(x_i)$, каждая из которых имеет сопряженную $\rho_i^*(x_i)$. Тогда для любой непрерывно дифференцируемой функции $f(x_1, \ldots, x_n)$ справедливо неравенство*

$$\mathbf{D}(f) \leqslant \int_{R^n} \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \rho_i^*(x_i) \prod_{j \neq i} \rho_i(x_i)\, dx_1 \ldots dx_n,$$

при условии существования интеграла как повторного.

10.4. Сходимость

✓ *Последовательность случайных величин X_n сходится к с. в. X* ***по вероятности***, $X_n \xrightarrow{P} X$, *если для любого $\varepsilon > 0$*

$$\boxed{\mathbf{P}(|X_n - X| > \varepsilon) \to 0 \quad \text{при} \quad n \to \infty.}$$

✓ *Последовательность случайных величин X_n сходится к с. в. X* ***в среднеквадратическом***, $X_n \xrightarrow{\text{с. к.}} X$, *если*

$$\boxed{\mathbf{E}(X_n - X)^2 \to 0.}$$

✓ *Последовательность случайных величин* X_n *сходится к с. в.* X ***почти наверное*** (*синоним*: *«с вероятностью* 1»), $X_n \xrightarrow{\text{п. н.}} X$, *если*

$$\boxed{\mathbf{P}\{|X_k - X| < \varepsilon,\ k \geqslant n\} \to 1 \quad \text{при} \quad n \to \infty.}$$

Поскольку X_n есть функция $X_n(\omega)$, то $X_n \xrightarrow{\text{п. н.}} X$, если $X_n(\omega)$ сходится к $X(\omega)$ в обычном смысле почти для всех ω, за исключением ω-множества нулевой меры.

✓ *Последовательность с. в.* X_n *называется фундаментальной — по вероятности, в среднем, почти наверное, — если, соответственно,*

$$\mathbf{P}(|X_n - X_m| > \varepsilon) \to 0, \qquad \mathbf{E}\,(X_n - X_m)^2 \to 0, \qquad \mathbf{P}\{|X_k - X_l| < \varepsilon;\ k, l \geqslant n\} \to 1,$$

при $m, n \to \infty$ *и* $\varepsilon > 0$.

✓ **Признак сходимости Коши.** Для сходимости $X_n \to X$ в любом указанном выше смысле необходима и достаточна фундаментальность последовательности X_n в том же смысле.

✓ Сходимость по вероятности из перечисленных разновидностей самая слабая. Импликация «$\xrightarrow{\text{п. н.}}$» $\Rightarrow$ «$\xrightarrow{P}$» очевидна, а неравенство Чебышева обеспечивает «$\xrightarrow{\text{с. к.}}$» $\Rightarrow$ «$\xrightarrow{P}$». Обратное в обоих случаях неверно.

В итоге $\left\{ \boxed{\xrightarrow{\text{п. н.}}} \atop \boxed{\xrightarrow{\text{с. к.}}} \right\} \Rightarrow$ «$\xrightarrow{P}$». Других импликаций нет.

✓ **Сходимость по распределению.** *Последовательность случайных величин* X_n *сходится к с. в.* X *по распределению,* $X_n \xrightarrow{D} X$, *если последовательность соответствующих функций распределения* $F_n(x)$ *слабо сходится к функции распределения* $F(x)$.

Слабая сходимость $F_n(x) \xrightarrow{w} F(x)$ означает

$$\mathbf{E}\,\{\phi(X_n)\} \to \mathbf{E}\,\{\phi(X)\}$$

для любой *непрерывной и ограниченной* функции $\phi(x)$. Это равносильно *поточечной сходимости* $F_n(x) \to F(x)$ *в точках непрерывности* $F(x)$.

✓ Импликация «$\xrightarrow{P}$» $\Rightarrow$ «$\xrightarrow{D}$» очевидна. Обратное неверно.

✓ **Теорема.** *Сходимость по распределению* $X_n \xrightarrow{D} X$ *равносильна равномерной (на любом конечном промежутке) сходимости* $\varphi_n(\lambda) \to \varphi(\lambda)$ *характеристических функций.*

✓ **Сходимость матожиданий.** *Пусть* $X_n \xrightarrow{\text{п. н.}} X$ *и все* $|X_n| < Y$, *где с. в.* Y *имеет конечное матожидание. Тогда* X *тоже имеет конечное матожидание и* $\mathbf{E}\{X_n\} \to \mathbf{E}\{X\}$.

Пусть $X_n \xrightarrow{\text{с. к.}} X$ *и все* $|X_n| < \infty$. *Тогда* $\mathbf{E}\{|X|\} < \infty$ *и* $\mathbf{E}\{X_n\} \to \mathbf{E}\{X\}$.

✓ Последовательность X_n называется *равномерно интегрируемой*, если

$$\sup_n \int\limits_{|x|>M} |x|\, dF_n(x) \to 0 \quad \textit{при} \quad M \to \infty,$$

где $F_n(x)$ функция распределения X_n.

✓ *В условиях равномерной интегрируемости* X_n:

(i) $\sup\limits_n \mathbf{E}\{|X_n|\} < \infty$;

(ii) *из* $X_n \xrightarrow{D} X$ *следует существование* $\mathbf{E}\{X\}$ *и* $\mathbf{E}\{X_n\} \to \mathbf{E}\{X\}$.

✓ **Закон «нуля или единицы».** *Если* $X_1, X_2, \ldots$ — *независимые случайные величины, а событие* A *определяется поведением только бесконечно далекого хвоста последовательности* $X_1, X_2, \ldots$ *и не зависит от значений* $X_1, \ldots, X_n$ *при любом конечном* n, — *то*

$$\textit{либо } \mathbf{P}\{A\} = 0, \quad \textit{либо } \mathbf{P}\{A\} = 1.$$

✓ События, зависящие только от «хвоста», называют *остаточными*. Таковы, например, *события*: *сходимости ряда* $\sum\limits_{k=1}^{\infty} X_k$ *либо самой последовательности* X_k; *ограниченности верхнего предела* $\overline{\lim\limits_{k\to\infty}} X_k < \infty$ и т. п.

✓ **Теорема.** *Если* $X_1, X_2, \ldots$ — *независимые случайные величины с нулевыми матожиданиями, то для сходимости ряда* $\sum\limits_{k=1}^{\infty} X_k$ *почти наверное достаточно сходимости числового ряда*:

$$\sum_{k=1}^{\infty} \mathbf{D}\{X_k\} < \infty. \tag{10.1}$$

А если все X_k *ограничены,* $\mathbf{P}\{|X_k| < M\} = 1$, *то условие* (10.1) *и необходимо.*

✓ Специфика случайных рядов (в отличие от последовательностей общего вида) проявляется в следующем полезном факте.

Теорема. *Если* $X_1, X_2, \ldots$ — *независимые случайные величины, то для ряда* $\sum\limits_{k=1}^{\infty} X_k$ *понятия сходимости почти наверное, по вероятности и по распределению — эквивалентны.*

✓ **Центральная предельная теорема.** Слабую сходимость

$$\lim_{n\to\infty} \mathbf{P}\left\{ \frac{S_n - \mathbf{E}\, S_n}{\sqrt{\mathbf{D}\, S_n}} < x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2}\, ds = \Phi(x)$$

обеспечивает *условие Ляпунова*: для некоторого $\delta > 0$

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} \mathbf{E}\,|X_k - m_k|^{2+\delta} \to 0 \quad \text{при} \quad n \to \infty,$$

а также более свободное *условие Линдеберга*: для любого τ

$$\frac{1}{B_n^2} \int\limits_{|x-m_k| \geqslant \tau B_n} (x - m_k)^2\, dF_k(x) \to 0 \quad \text{при} \quad n \to \infty,$$

где $F_k(x)$ — функция распределения X_k.

10.5. Марковские процессы

✓ *Марковским процессом* называют *последовательность случайных величин (векторов) $X_1, \ldots, X_n, \ldots$, в которой «будущее» $X_{t>n}$ определяется только величиной X_n и не зависит от предыстории $X_1, \ldots, X_{n-1}$.*

✓ Марковский процесс с дискретным временем и *счетным* пространством состояний называют *марковской цепью*. Как правило, подразумевается следующая модель. Состояния пронумерованы. Система (частица), находясь в k-й момент времени в j-м состоянии в $(k+1)$-й момент попадает в i-е состояние с вероятностью P_{ij}, и тогда при распределении частицы по состояниям с вероятностями p_j^k в следующий момент получается распределение

$$p_i^{k+1} = \sum_j P_{ij} p_j^k,$$

или в векторном виде $\mathbf{p}^{k+1} = P\mathbf{p}^k$, где $P = [P_{ij}]$ называют *матрицей переходных вероятностей*.

Динамика распределений $\mathbf{p}^k$ определяется итерациями матрицы P:

$$\mathbf{p}^{k+m} = P^m \mathbf{p}^k.$$

Стационарные распределения $\mathbf{p}^*$ оказываются собственными векторами матрицы P,

$$\mathbf{p}^* = P\mathbf{p}^*,$$

а сходимость $\mathbf{p}^k \to \mathbf{p}^*$ — одним из центральных вопросов.

✓ Отличительной особенностью матриц переходных вероятностей является условие $\sum_i P_{ij} = 1$, т. е. все столбцовые суммы единичны. Такие матрицы $P \geqslant 0$ называют *стохастическими*.

• Собственный вектор $\mathbf{p}^* \geqslant 0$, отвечающий собственному значению $\lambda = 1$, у стохастической матрицы существует всегда, т. е. всегда существует стационарное распределение $\mathbf{p}^* = P\mathbf{p}^*$.

• Если матрица P строго положительна (все $P_{ij} > 0$) или же $P^k > 0$ при некотором k, то все стационарные вероятности $p_j^* > 0$, причем итерации $\mathbf{p}^k$ сходятся к $\mathbf{p}^* > 0$, а итерации $P^k \to P_\infty$, где у P_∞ все столбцы одинаковы и равны $\mathbf{p}^*$. Процесс в этом случае называют *эргодическим*.

• Условие «$P^k > 0$ при некотором k» необходимо и достаточно для *примитивности* стохастической матрицы, т. е. для того, чтобы спектр P, за исключением ведущего собственного значения $\lambda = 1$, лежал строго внутри единичного круга. В случае *импримитивной* (не примитивной) матрицы P предел $\mathbf{p}^k$ может не существовать. Но предел имеют средневзвешенные суммы,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} P^k = P_\infty.$$

✓ Матрица P называется *разложимой* (*неразложимой*), если одинаковой перестановкой строк и столбцов она приводится (не приводится) к виду

$$\begin{bmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{bmatrix},$$

где P_{11} и P_{22} квадратные матрицы.

✓ *Если матрица P неразложима, то $\lambda(P) = 1$ является ведущим собственным значением P алгебраической кратности* 1, *которому отвечает строго положительный собственный вектор. Других положительных собственных значений и векторов у P нет.*

10.6. Случайные функции и процессы

✓ *Случайной функцией* называют функцию двух переменных $X(t, \omega)$, где ω — точка вероятностного пространства Ω, на котором задана та или иная вероятностная мера. Зависимость от случая реализуется при этом каждый раз наступлением исхода $\omega_0 \in \Omega$, при котором фактическое течение процесса описывается *траекторией* $X(t)$, которую называют также *реализацией процесса* или *выборочной функцией*.

✓ Плотность $\rho(x, t)$ случайной функции $X(t)$ определяет распределение значений $X(t)$ в момент t.

Для с. ф естественным образом определяются: *матожидание*

$$m_x(t) = \mathrm{E}\{X(t)\} = \int_{-\infty}^{\infty} x\rho(x, t)\, dx$$

и *корреляционная функция*

$$R_{xx}(t, s) = \mathbf{E}\{[X(t) - m_x(t)][X(s) - m_x(s)]\},$$

которая при $t = s$ превращается в *дисперсию*

$$D_x(t) = R_{xx}(t, t) = \mathbf{E}\{[X(t) - m_x(t)]^2\}.$$

✓ Случайный процесс $X(t)$ *стационарен*, если его характеристики не меняются при сдвиге по оси времени.

При независимости от сдвига по оси времени n-мерной плотности распределения с. ф. $X(t)$ называют *стационарной в узком смысле.*

Менее жесткий вариант: независимость от сдвига по оси времени условного матожидания и корреляционной функции. В этом случае с. ф. $X(t)$ называют *стационарной в широком смысле.*

В том и другом случае матожидание и дисперсия не зависят от времени, а корреляция $R_{xx}(t, s)$ зависит только от разности $t - s$.

✓ С. ф. $X(t)$ называют *эргодичной* (по отношению к матожиданию) при равенстве среднего значения X по ансамблю и — среднего по времени. Для стационарного процесса это означает

$$\lim_{T\to\infty} \mathbf{E}\left\{\left[\frac{1}{T}\int_{t_0}^{t_0+T} X(t)\,dt - m_x\right]^2\right\} = 0,$$

где t_0 — произвольный момент времени, а $m_x = \mathbf{E}\{X(t)\}$.

Об эргодичности можно говорить по отношению к любой функции

$$Y = \varphi[X(t_1), \ldots, X(t_n)].$$

В частности, — по отношению к корреляционной функции, отталкиваясь от

$$Y(t, s) = [X(t) - m_x][X(s) - m_x].$$

Эргодическое свойство позволяет экспериментально определять матожидание любой стационарной функции $Y(t) = \varphi[X(t)]$ не по множеству реализаций, а по данным одной реализации на достаточно большом промежутке времени T.

Эргодичность стационарной функции по отношению к матожиданию обеспечивает условие

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \left(1 - \frac{\tau}{T}\right) R_{xx}(\tau)\,d\tau = 0.$$

✓ Преобразование Фурье $\widehat{R}(\omega)$ корреляционной функции стационарного процесса,

$$\widehat{R}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R(\tau)e^{-i\omega\tau}\,d\tau \quad \Leftrightarrow \quad R(\tau) = \int_{-\infty}^{\infty} \widehat{R}(\omega)e^{i\omega\tau}\,d\omega,$$

называют *спектральной плотностью* сигнала $X(t)$.

Взаимосвязь спектра корреляционной функции со спектром самого сигнала $X(t)$ дает соотношение

$$\widehat{R}(\omega) = \lim_{T\to\infty} \frac{2\pi}{T} \mathrm{E}\{|\widehat{A}_T(\omega)|^2\},$$

где $\widehat{A}_T(\omega)$ — преобразование Фурье сигнала $A_T(t) = X_T(t) - m_x$, совпадающего с $X(t) - m_x$ на промежутке $t \in [-T/2, T/2]$ и равного нулю вне этого промежутка.

Широкое распространение имеет энергетическое соотношение

$$D_{xx} = \sigma_x^2 = R_{xx}(0) = \int_{-\infty}^{\infty} \widehat{R}_{xx}(\omega)\, d\omega,$$

увязывающее среднюю мощность случайного сигнала с его спектральной плотностью.

✓ Стационарный случайный сигнал $X(t)$ с постоянной спектральной плотностью

$$\widehat{R}_{xx}(\omega) \equiv G$$

во всем диапазоне частот от нуля до бесконечности — называют *белым шумом*.

Обратное преобразование Фурье приводит в этом случае к дельтаобразной корреляционной функции

$$R_{xx}(\tau) = G \int_{-\infty}^{\infty} e^{i\omega\tau}\, d\omega = 2\pi G\delta(\tau).$$

✓ Случайная функция $X(t)$ называется *процессом с независимыми приращениями*, если для любых $t_0 < t_1 < \ldots < t_n$ случайные величины $X(t_1) - X(t_0), \ldots X(t_n) - X(t_{n-1})$ независимы.

Процесс считается *однородным*, если распределение

$$X(t) - X(s)$$

определяется только разностью $t - s$.

Однородный процесс $X(t)$ с независимыми приращениями называют *броуновским движением*, или *винеровским процессом*, если все $X(t_k) - X(t_{k-1})$ распределены нормально со средним 0 и дисперсией $|t_k - t_{k-1}|$.

✓ Дифференцирование случайной функции перестановочно с операцией математического ожидания. Формула для вычисления корреляционной функции производной $Y(t) = X'(t)$,

$$\boxed{R_{yy}(t, s) = \frac{\partial^2 R_{xx}(t, s)}{\partial t \partial s},}$$

легко получается предельным переходом.

Спектральная плотность производной сигнала $Y(t) = X'(t)$ равна

$$\boxed{\widehat{R}_{yy}(\omega) = \omega^2 \widehat{R}_{xx}(\omega).}$$

✓ Понимание метаморфоз, которые происходят со случайными сигналами при их интегрировании и дифференцировании, играет важную роль в изучении динамичееких систем, описываемых дифференциальными уравнениями.

В отличие от детерминированных систем, преобразование Фурье выходного сигнала, рáвно как и сам сигнал, — для понимания ситуации ничего особенно не дают. Здесь важны не беспорядочные флуктуации, а вероятностные характеристики сигнала, определяемые преобразованием спектра:

$$\boxed{\widehat{R}_{yy}(\omega) = |W(i\omega)|^2 \widehat{R}_{xx}(\omega),}$$

где W — передаточная функция линейной системы.

10.7. Теория информации

✓ *Энтропия* (*неопределенность*) случайной величины, принимающей n различных значений с вероятностями $p_1, \ldots, p_n$, определяется как

$$\boxed{H(p_1, \ldots, p_n) = -\sum p_i \log_2 p_i.}$$

При этом действует соглашение $\boxed{0 \cdot \log 0 = 0}$. Двойка в основании логарифмов обычно опускается, а единица измерения называется *битом*. Таким образом, бит соответствует неопределенности выбора из двух равновероятных возможностей (то ли нуль, то ли единица — например).

✓ Важную роль играют свойства энтропии при рассмотрении объединенных систем. Пусть $\{x_1, \ldots, x_n\}$ и $\{y_1, \ldots, y_n\}$ — возможные состояния случайных величин X и Y либо двух систем X и Y. Состояния вектора $\{X, Y\}$ представляют собой комбинации пар x_i и y_i. Энтропия $\{X, Y\}$ по определению равна

$$H(X, Y) = -\sum_{i,j} p_{ij} \ln p_{ij},$$

где $p_{ij} = p(x_i, y_j) = \mathrm{P}\{X = x_i,\ Y = y_j\}$.

Если системы X и Y *независимы*, то $p_{ij} = p_i p_j$, и

$$H(X, Y) = H(X) + H(Y).$$

Если же системы *зависимы*, то $p(x_i, y_j) = p(x_i)p(y_j|x_i)$, и

$$H(X, Y) = H(X) + H(Y|X),$$

где

$$H(Y|X) = \sum_i p(x_i)H(Y|x_i)$$

называют *полной условной энтропией*, а

$$H(Y|x_i) = -\sum_j p(y_j|x_i)\log_2 p(y_j|x_i)$$

условной энтропией Y при условии $X = x_i$.

В обоих случаях говорят об *аддитивности энтропии.*

✓ Простейшие свойства энтропии:

- *Энтропия всегда неотрицательна и достигает максимума в случае равновероятных возможностей.*
- *Пусть* $\sum p_k = \sum q_k = 1$, *т. е.* p_k *и* q_k — *два распределения, причем все* $q_k > 0$. *Тогда*

$$\boxed{\sum_k p_k \log p_k \geqslant \sum_k p_k \log q_k.}$$

- *Условная энтропия всегда меньше или равна безусловной*

$$H(Y|X) \leqslant H(Y),$$

причем при добавлении условий энтропия не увеличивается.

✓ Пусть $H(A)$ — энтропия исхода некоторого опыта A. Если опыт B содержит какие-то сведения относительно A, то после проведения B неопределенность A уменьшается до условной энтропии $H(A|B)$. Разность

$$\boxed{I(A, B) = H(A) - H(A|B),}$$

по определению, есть *количество информации*, содержащееся в B относительно A.

✓ **Энтропия источника.** Если источник информации потенциально может передать i-й символ (алфавита) с вероятностью p_i, то матожидание передаваемой информации (на один символ) при длительной работе источника — равно энтропии источника:

$$I = -\sum p_i \log_2 p_i.$$

В итоге ясно, что информация и энтропия — это две стороны одного явления. Сколько поступает информации — настолько убывает энтропия (неопределенность). Чем больше энтропия источника, тем больше информации при получении его сигналов. Источник, способный генерировать единственный сигнал, никакой информации не производит. Источник, передающий только два сигнала «ноль/один», имеет единичную интенсивность (один бит на сигнал). Но при большой частоте способен производить много бит в единицу времени.

✓ **Пропускная способность канала.** Канал связи в схеме

$$\boxed{\textit{источник}} \Rightarrow \boxed{\textit{канал связи}} \Rightarrow \boxed{\textit{приемник}},$$

так или иначе, ограничивает скорость передачи информации. В простейшем и широко распространенном случае, когда символов (сигналов) всего два и их длительности одинаковы, *пропускная способность* C измеряется числом символов, способных пройти по каналу за одну секунду.

В общем случае C — это максимальная информация, которая может быть передана по каналу за одну секунду. Если, например, алфавит состоит из n букв и канал способен пропускать N букв в секунду (в точности или в среднем), то $C = N \log_2 n$.

✓ **Частотная интерпретация.** Пусть источник генерирует i-й символ с вероятностью p_i и символы в сообщении длины N независимы. При достаточно большом N количество символов i-го вида в сообщении с большой точностью равно Np_i. Это дает вероятность сообщения

$$p = p_1^{Np_1} \dots p_n^{Np_n},$$

т. е.

$$\log p = N \sum p_i \log_2 p_i \quad \Rightarrow \quad \boxed{p = 2^{-NH}}.$$

Иными словами, вероятности всех достаточно длинных сообщений равны $p = 2^{-NH}$, а поскольку эти сообщения еще и независимы, то их количество $K = 1/p$, т. е.

$$\boxed{K = 2^{NH}.}$$

Таким образом, энтропия по правилу $K = 2^{NH}$ определяет, например, *количество текстов, в которых буквы встречаются с «правильной» частотой.*

При вероятностной (не частотной) трактовке это означает следующее. С какими бы вероятностями p_i источник ни генерировал символы — принципиально возможны все n^N сообщений Q длины N, но их вероятности $p(Q)$ различны.

Тогда при любом $\varepsilon > 0$

$$\lim_{N\to\infty} \sum_{|p(Q)-2^{-NH}|>\varepsilon} p(Q) = 0,$$

т. е. сумма вероятностей всех сообщений, вероятности которых отличаются от 2^{-NH} *более чем на* ε, *— стремится к нулю (сколь угодно мала при большом* N*).*

Соответственно, вероятности сообщений

$$p(Q) \in (2^{-NH} - \varepsilon,\ 2^{-NH} + \varepsilon)$$

в сумме стремятся к 1. Поэтому при больших N можно считать, что «наблюдаемых» сообщений (последовательностей, текстов) имеется как бы ровно 2^{NH}. Остальными можно пренебречь — их суммарная вероятность близка к нулю.

✓ **Оптимальное кодирование.** Естественное соображение при кодировании: часто встречающимся символам и словам исходного сообщения ставить в соответствие короткие «01»-комбинации, редко встречающимся — длинные. Если в результате символы 0 и 1 будут встречаться одинаково часто, — это будет оптимальным кодом.

Оптимальную «игру» на длине кодовых комбинаций реализует *код Шеннона—Фано*. Буквы алфавита упорядочиваются по убыванию частоты (вероятности) p_i появления в тексте, после чего разбиваются на две группы. К первой — относят первые k букв — так, чтобы

$$\sum_{i=1}^{k} p_i \approx \sum_{i=k}^{n} p_i \approx \frac{1}{2},$$

после чего первой группе символов ставится в соответствие 0, второй — 1, и это определяет первый разряд кодового числа. Далее каждая группа снова делится на две приблизительно равновероятные подгруппы; первой подгруппе ставится в соответствие 0, второй — 1 и т. д. Группы с малым количеством букв быстро исчерпываются — и эти буквы в результате получают короткие коды. Легко убедиться, что в итоге кодовая запись достаточно длинного сообщения будет содержать приблизительно одинаковое количество нулей и единиц, т. е. при любой частотности исходных символов частоты нулей и единиц двоичных кодов оказываются $\approx$ равны друг другу.

Информационная сторона оптимального кодирования в общем виде выглядит так. Равновероятные сообщения в количестве $K = 2^{NH}$ могут быть пронумерованы в двоичной записи, для чего потребуется минимальное число разрядов $\log_2 K = NH$. Это и будет оптимальным двоичным кодом.

В рамках вероятностной модели возможны все n^N сообщений длины N (а не только $K = 2^{NH}$), но при больших N можно считать, что «наблюдаемых» сообщений имеется как бы ровно 2^{NH}. Остальными можно пренебречь — их суммарная вероятность близка к нулю. Поэтому маловероятные сообщения можно кодировать достаточно длинными «01-последовательностями». Из-за их маловероятности это в среднем почти не будет сказываться на скорости передачи информации.

✓ **Канал с шумом.** При наличии шума в канале связи,

$$\boxed{\text{вход } X} \Rightarrow \overset{\Downarrow \xi}{\boxed{\text{канал связи}}} \Rightarrow \boxed{\text{выход } Y = f(X, \xi)}$$

выходной сигнал $Y = f(X, \xi)$ зависит от входа X и шума ξ.

Если шум искажает в среднем 1 % символов, то о любом принятом символе нельзя сказать наверняка, правилен он или нет. Максимум возможного — при

независимой генерации букв — утверждать их правильность с вероятностью 0,99. Но если речь идет о передаче осмысленного текста, то сообщение при 1 % ошибок можно восстановить (по словарю) с высокой степенью надежности. Понятно, что это возможно благодаря избыточности языка. В общем случае проблема заключается в том, чтобы подобную избыточность обеспечить при кодировании.

В нешумящем канале $H(X|Y) = 0$, т. е. принятый сигнал однозначно определяет переданный. В общем случае условная энтропия $H(X|Y)$ служит показателем того, насколько шумит канал. При вероятности ошибки 0,01 в случае равновероятной передачи источником двоичных символов

$$H(X|Y) = -\frac{1}{100}\log\frac{1}{100} - \frac{99}{100}\log\frac{99}{100} \approx 0{,}08 \quad \textit{бит на символ.}$$

Поэтому при передаче по каналу 100 символов в секунду скорость передачи информации равна $100 - 8 = 92$ бита в секунду. Ошибочно принимается лишь один бит из ста, но «потери» равны 8 битам из-за того, что неясно, какой символ принят неверно.

Пропускная способность канала с шумом, по определению Шеннона, — это максимальная скорость прохождения информации

$$\boxed{C = \max[H(X) - H(X|Y)]} \quad (\textit{бит в секунду}),$$

где максимум берется по всем возможным источникам информации, а энтропия H измеряется в ***битах в секунду***.

На первый взгляд, это сильно отличается от ситуации канала без шума, где под C обычно мыслится максимально возможное число проходящих импульсов. Но это не совсем так. Во-первых, система передачи может быть не двоичной. Во-вторых, сама передача символов по каналу бывает малоэффективна — символов много, информации мало. Поэтому аккуратное определение пропускной способности канала без шума в точности совпадает с данным выше определением, при условии $H(X|Y) = 0$. При этом ясно, что в ситуации $H > C$ передача информации без потерь невозможна.

В примере с искажением 1 % двоичных символов, если канал физически способен пропускать 100 бит/с, — его пропускная способность равна 92 бит/с. Информационные потери 8 бит приходятся на $H(X|Y)$, т. е. на шум.

Теоремы Шеннона. Допустим, что помимо основного — есть дополнительный корректирующий канал.

Если корректирующий канал имеет пропускную способность не меньше $H(X|Y)$, то при надлежащей кодировке возможен практически безошибочный прием сообщений (с точностью до сколь угодно малой доли ошибок).

Пусть H бит/с — энтропия источника, а C — пропускная способность канала с шумом. Если $H \leqslant C$, то при надлежащем кодировании возможен практически безошибочный прием сообщений (с точностью до сколь угодно малой доли ошибок).

✓ **Энтропия непрерывного распределения** $\rho(x)$ определяется как

$$H = -\int_{-\infty}^{\infty} \rho(x) \log \rho(x)\, dx.$$

Если X — случайный вектор, энтропия вычисляется аналогично с той лишь разницей, что интегрирование ведется по всему пространству.

Свойства энтропии непрерывных распределений в основном подобны свойствам энтропии дискретных распределений. Максимум H на ограниченной области достигается при равномерной плотности, а максимум при заданной дисперсии — приводит к нормальному закону.

✓ Любой непрерывный сигнал с ограниченным спектром, в силу теоремы отсчетов (теоремы Котельникова), может быть представлен в виде

$$x(t) = \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2W}\right) \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)},$$

т. е. определяется значениями $x(t)$ в дискретном ряде точек, расположенных с интервалом времени $\Delta t = 1/(2W)$, где W — полоса пропускания частот (ширина спектра). Это позволяет свести изучение передачи непрерывных сигналов к дискретному случаю.

Пусть

$$Y(t) = X(t) + N(t),$$

где $X(t)$ — передаваемый сигнал, $Y(t)$ — принимаемый, $N(t)$ — белый шум мощности D_N. В силу независимости $X(t)$ и $N(t)$, мощность (дисперсия) сигнала на выходе равна $D_Y = D_X + D_N$.

Вычисление показывает, что пропускная способность канала в данном случае,

$$C = W \log_2\left(1 + \frac{D_X}{D_N}\right),$$

определяется полосой пропускания W и отношением *сигнал/шум*, D_X/D_N.

10.8. Статистика

✓ Набор независимых случайных величин $X_1, \ldots, X_n$, каждая из которых распределена так же, как изучаемая с. в. X, — называют *случайной выборкой* объема n, а любую функцию $\Theta_n = \Theta_n(X_1, \ldots, X_n)$ — *статистической характеристикой* (с. х.), или *статистикой*. Определению обычно подлежат вероятности тех

или иных событий, матожидания, дисперсии, корреляции и другие характеристики с. в. на базе с. х.

✓ При оценке неизвестного параметра θ, характеризующего с. в. X, оценка θ на основе Θ_n называется — *состоятельной*, если $\Theta_n \xrightarrow{P} \theta$ при $n \to \infty$, и — *смещенной/несмещенной*, если матожидание $\mathbf{E}\{\Theta_n\}$ при любом n *равно/не равно* θ.

✓ Промежуток, которому принадлежит оцениваемый параметр θ с вероятностью $\geqslant \delta$, называют *доверительным интервалом*, δ — *коэффициентом доверия*, а $1 - \delta$ — *уровнем значимости*.

✓ В случае существования у с. в. X первых двух моментов выборочное среднее

$$\widehat{X}_n = \frac{X_1 + \ldots + X_n}{n},$$

в силу $\mathbf{E}\{\widehat{X}_n\} = m_x$, является несмещенной оценкой. Плюс к тому,

$$\mathbf{D}\{\widehat{X}_n\} = \frac{D_x}{n},$$

что обеспечивает $\widehat{X}_n \xrightarrow{\text{с. к.}} X$, и тем более, $\widehat{X}_n \xrightarrow{P} X$.

Однако несмещенной оценкой дисперсии является

$$\widehat{D}'_n = \frac{(X_1 - \widehat{X}_n)^2 + \ldots + (X_n - \widehat{X}_n)^2}{n - 1},$$

где в знаменателе стоит $n - 1$ вместо интуитивно ожидаемого n.

✓ *Хи-квадрат* распределение имеет плотность $\rho(x) = 0$ при $x \leqslant 0$ и

$$\rho(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} \quad \text{при} \quad x > 0,$$

где Γ — гамма-функция [2], а целочисленный параметр n называют *числом степеней свободы*.

Так распределен квадрат вектора $\chi = \{X_1, \ldots, X_n\}$,

$$\chi^2 = X_1^2 + \ldots + X_n^2,$$

с нормальными координатами X_k, имеющими нулевые матожидания и единичные дисперсии.

Распределение Стьюдента (t-распределение) имеет случайная величина

$$t = \frac{\sqrt{n}X}{\sqrt{\chi^2}},$$

---

[2] $\Gamma(\nu) = \int_0^\infty x^{\nu-1} e^{-x}\, dx.$

где n *число степеней свободы*, с. в. X имеет нормальное распределение $\mathcal{N}(0, 1)$, а χ^2 распределена по закону «хи-квадрат».

✓ Для независимых и одинаково распределенных с. в. $X_1, \ldots, X_n$ с плотностью $\rho_\theta(x)$ *метод максимального правдоподобия* Фишера заключается в максимизации совместной плотности

$$\rho_\theta(x_1, \ldots, x_n) = \prod_{k=1}^{n} \rho_\theta(x_k)$$

распределения $X_1, \ldots, X_n$ при полученных реализациях $x_1, \ldots, x_n$. Функция $\Theta_n = \Theta_n(x_1, \ldots, x_n)$, обеспечивающая такой максимум, называется *оценкой максимального правдоподобия*.

Неравенство Рао—Крамера:

$$\mathbf{E}\{(\Theta_n - \theta)^2\} \geqslant \frac{1}{I_n(\theta)},$$

где

$$I_n(\theta) = \mathbf{E}\left\{\frac{\partial}{\partial\theta} \ln \rho_\theta(X_1, \ldots, X_n)\right\}$$

— *количество информации по Фишеру.*

Сокращения и обозначения

ТВ — теория вероятностей

с. в. — случайная величина

с. ф. — случайная функция

с. х. — статистическая характеристика

с. к. — среднеквадратический(ая, ое)

п. ф. — производящая функция

х. ф. — характеристическая функция

п. н. — почти наверное

б. ч. р. — бесконечное число раз

◄ и ► — начало и конец рассуждения, темы, доказательства

(**?**) — предлагает проверить или доказать утверждение в качестве упражнения, либо довести рассуждение до «логической точки»

(**!**) — предлагает обратить внимание

$\int$ — обозначает интегрирование по области определения функции, стоящей под интегралом, чаще всего: $\int\limits_{-\infty}^{\infty}$

$\mathbf{P}(A)$ — вероятность события A

$\mathbf{E}(X)$ — математическое ожидание случайной величины X

$\mathbf{D}(X) = \mathbf{E}[X - \mathbf{E}(X)]^2$ — дисперсия случайной величины X

$\mu_r = \mathbf{E}\,[X - \mathbf{E}\,(X)]^r$ — центральный момент r-го порядка

$\sigma_x^2 = \sqrt{\mathbf{D}\,(X)}$ — среднеквадратическая ошибка

$\mathcal{N}(m_x, \sigma_x^2)$ — нормальное распределение с матожиданием m_x и дисперсией σ_x^2

Ω — пространство элементарных событий

$A \Rightarrow B$ — из A следует B

$x \in X$ — x принадлежит X

$X \cup Y$, $X \cap Y$, $X \backslash Y$ — объединение, пересечение и разность множеств X и Y

$X \triangle Y = (X \backslash Y) \cup (Y \backslash X)$ — симметрическая разность множеств X и Y

$X \subset Y$ — X подмножество Y, в том числе имеется в виду возможность $X \subseteq Y$, т. е. между $X \subset Y$ и $X \subseteq Y$ различия не делается

$\varnothing$ — пустое множество

i — мнимая единица, $i^2 = -1$

$z = x + iy$ — комплексное число, $z = r(\cos\varphi + i\sin\varphi)$ — его тригонометрическая запись, $x = \operatorname{Re} z$ — действительная часть, $y = \operatorname{Im} z$ — мнимая; $\overline{z} = z^* = x - iy$ — комплексно сопряженное число

(x, y) либо $\langle x, y\rangle$ — скалярное произведение векторов x и y; в общем случае комплексных векторов

$$\langle x, y\rangle = x_1 y_1^* + \ldots + x_n y_n^*;$$

для скалярного произведения используются также эквивалентные обозначения $x \cdot y$ и xy

$|A| = \det A$ — определитель (детерминант) матрицы A

$\rho(A)$ — спектральный радиус матрицы A

$\dfrac{df(t)}{dt} = f'(t)$ — производная $f(t)$

$\dfrac{\partial u}{\partial x}$ — частная производная функции u по переменной x; эквивалентное обозначение u'_x

$\nabla f(x)$ — градиент функции $f(x)$

Литература

1. *Беккенбах Э., Беллман Р.* Неравенства. М., 2004.
2. *Билингслей П.* Эргодическая теория и информация. М.: Мир, 1969.
3. *Боровков А. А.* Теория вероятностей. М.: УРСС, 2003.
4. *Босс В.* Интуиция и математика. М., 2003.
5. *Босс В.* Лекции по математике. М.: УРСС, 2004–2005.
6. *Гихман И. И., Скороход А. В.* Введение в теорию случайных процессов. М.: Наука, 1965.
7. *Данцер Л., Грюнбаум Б., Кли В.* Теорема Хелли и ее применения. М.: Мир, 1968.
8. *Дуб Д.* Вероятностные процессы. М.: ИЛ, 1956.
9. *Дюге Д.* Теоретическая и прикладная статистика. М.: Наука, 1972.
10. *Золотарев В. М.* Современная теория суммирования независимых случайных величин. М.: Наука, 1986.
11. *Кац М.* Статистическая независимость в теории вероятностей, анализе и теории чисел. М.: ИЛ, 1963.
12. *Кац М.* Вероятность и смежные вопросы в физике. М.: УРСС, 2003.
13. *Колмогоров А. Н.* Основные понятия теории вероятностей. М.: Наука, 1974.
14. *Колмогоров А. Н., Фомин С. В.* Элементы теории функций и функционального анализа. М.: Наука, 1972.
15. *Ламперти Дж.* Вероятность. М.: Наука, 1973.
16. *Липцер Р. Ш., Ширяев А. Н.* Теория мартингалов. М.: Наука, 1986.
17. *Неве Ж.* Математические основы теории вероятностей. М.: Мир, 1969.
18. *Прохоров Ю. В., Розанов Ю. А.* Теория вероятностей. СМБ. М.: Наука, 1973.
19. *Прохоров Ю. В., Ушаков В. Г., Ушаков Н. Г.* Задачи по теории вероятностей. М.: Наука, 1986.
20. *Пугачев В. С.* Теория вероятностей и математическая статистика. М.: Наука, 1979.
21. *Розанов Ю. А.* Теория вероятностей, случайные процессы и математическая статистика. М.: Наука, 1985.

22. *Секей Г.* Парадоксы в теории вероятностей и математической статистике. М.: Мир, 1990.
23. *Спицер Ф.* Принципы случайного блуждания. М.: Мир, 1969.
24. *Стэнли Р.* Перечислительная комбинаторика. М.: Мир, 1990.
25. *Уиттл П.* Вероятность. М.: Наука, 1982.
26. *Феллер В.* Введение в теорию вероятностей и ее приложения. М.: Мир, 1967.
27. *Халмош П.* Теория меры. М.: ИЛ, 1953.
28. *Харрис Т.* Теория ветвящихся случайных процессов. М.: Мир, 1966.
29. *Хида Т.* Броуновское движение. М.: Наука, 1987.
30. *Шеннон К.* Работы по теории информации и кибернетике. М.: ИЛ, 1963.
31. *Ширяев А. Н.* Вероятность. М.: Наука, 1980.

Предметный указатель

ББК 22.171я73

Босс В.

Лекции по математике. Т. 4: **Вероятность, информация, статистика.**
М.: КомКнига, 2005. — 216 с.

ISBN 5-484-00168-4

Книга отличается краткостью и прозрачностью изложения. Объяснения даются «человеческим языком» — лаконично и доходчиво. Значительное внимание уделяется мотивации результатов. Помимо классических разделов теории вероятностей освещается ряд новых направлений: нелинейный закон больших чисел, асимптотическое агрегирование. Изложение сопровождается большим количеством примеров и парадоксов, способствующих рельефному восприятию материала. Затрагиваются многие прикладные области: управление запасами, биржевые игры, массовое обслуживание, страховое дело, стохастическая аппроксимация, обработка статистики. Несмотря на краткость, достаточно полно излагается теория информации с ответвлениями «энтропийно термодинамического» характера. Охват тематики достаточно широкий, но изложение построено так, что можно ограничиться любым желаемым срезом содержания. Книга легко читается.

Для студентов, преподавателей, инженеров и научных работников.

Издательство «КомКнига». 117312, г. Москва, пр-т 60-летия Октября, 9.
Подписано к печати 21.06.2005 г. Формат 60×90/16. Печ. л. 13,5. Зак. № 130.
Отпечатано в ООО «ЛЕНАНД». 117312, г. Москва, пр-т 60-летия Октября, д. 11А, стр. 11.

ISBN 5-484-00168-4

НАУЧНАЯ И УЧЕБНАЯ ЛИТЕРАТУРА
E-mail: URSS@URSS.ru
Каталог изданий в Интернете:
http://URSS.ru
Тел./факс: 7 (095) 135-42-16
URSS Тел./факс: 7 (095) 135-42-46

3299 ID 29480

Проект издания 20 томов «Лекций по математике» В. Босса обретает краски.

Автор наращивает обороты и поднимает планку все выше.

Читательская аудитория ширится.

В условиях информационного наводнения инструменты вчерашнего дня перестают работать.
Поэтому учить надо как-то иначе.
«Лекции» дают пример.
Плохой ли, хороший – покажет время.
Но в любом случае, это продукт нового поколения.
Те же «колеса», тот же «руль», та же математическая суть, – но по-другому.

В. Босс

Из отзывов читателей:

Чтобы усвоить предмет, надо освободить его от деталей, обнажить центральные конструкции, понять, как до теорем можно было додуматься. Это тяжелая работа, на которую не всегда хватает сил и времени. В «Лекциях» такая работа проделывается автором.

Популярность книг В. Босса среди преподавателей легко объяснима. Дается то, чего недостает. Общая картина, мотивация, взаимосвязи. И самое главное — легкость вхождения в любую тему.

Содержание продумано и хорошо увязано. Громоздкие доказательства ужаты до нескольких строчек. Виртуозное владение языком. Что касается замысла изложить всю математику в 20 томах, с трудом верится, что это по силам одному человеку.

Лекции В. Босса — замечательные математические книги. Как учебные пособия, они не всегда отвечают канонам преподавания, но студентам это почему-то нравится.

НАУЧНАЯ И УЧЕБНАЯ ЛИТЕРАТУРА
E-mail: URSS@URSS.ru
Каталог изданий в Интернете:
http://URSS.ru
Тел./факс: 7 (095) 135–42–16
URSS Тел./факс: 7 (095) 135–42–46

3299 ID 29480
9 785484 001682 >

Отзывы о настоящем издании, а также обнаруженные опечатки присылайте по адресу **URSS@URSS.ru**. Ваши замечания и предложения будут учтены и отражены на web-странице этой книги в нашем интернет-магазине **http://URSS.ru**