ЗАМЕЧАНИЯ ОБ УПРАЖНЕНИЯХ

Упражнения, помещенные в книгах настоящей серии, предназначены как для самостоятельной проработки, так и для семинарских занятий. Трудно, если не невозможно изучить предмет, только читая теорию и не применяя полученную информацию для решения специальных задач и тем самым не заставляя себя обдумывать то, что было прочитано. Кроме того, мы лучше всего заучиваем то, что сами открываем для себя. Поэтому упражнения образуют важную часть данной работы; были предприняты определенные попытки, чтобы отобрать упражнения, в которых бы содержалось как можно больше информации и которые было бы интересно решать.

Во многих книгах легкие упражнения даются вперемешку с исключительно трудными. Зачастую это очень неудобно, так как перед тем, как приступать к решению задачи, читатель обязательно должен представлять себе, сколько времени уйдет у него на это решение (иначе он может разве только просмотреть все задачи). Классическим примером здесь является книга Ричарда Беллмана "Динамическое программирование"; это важная пионерская работа, в которой в конце каждой главы под рубрикой "Упражнения и исследовательские проблемы" дается целый ряд задач, где наряду с глубокими еще нерешенными проблемами встречаются исключительно тривиальные вопросы. Говорят, что однажды кто-то спросил д-ра Беллмана, как отличить упражнения от исследовательских проблем, и тот ответил: "Если вы можете решить задачу, это—упражнение; в противном случае это—проблема".

Можно привести много доводов в пользу того, что в книге типа этой должны быть как исследовательские проблемы, так и очень простые упражнения, и для того чтобы читателю не приходилось ломать голову над тем, какая задача легкая, а какая трудная, мы ввели "оценки", которые указывают степень трудности каждого упражнения. Эти оценки имеют следующее значение:

Оценка	**Объяснение**
00	Чрезвычайно легкое упражнение, на которое можно ответить сразу же, если понят материал текста, и которое почти всегда можно решить "в уме".
10	Простая задача, которая заставляет задуматься над прочитанным материалом, но не представляет никаких особых трудностей. На решение такой задачи требуется не больше одной минуты; в процессе решения могут понадобиться карандаш и бумага.
20	Задача средней трудности, позволяющая проверить, насколько хорошо понят текст. На то чтобы дать исчерпывающий ответ, требуется примерно 15–20 минут.
30	Задача умеренной трудности и/или сложности, для удовлетворительного решения которой требуется больше двух часов.
40	Очень трудная или трудоемкая задача, которую, вероятно, следует включить в план практических занятий. Предполагается, что студент может решить такую задачу, но для этого ему потребуется значительный отрезок времени; задача решается нетривиальным образом.
50	Исследовательская проблема, которая (насколько это было известно автору в момент написания) еще не получила удовлетворительного решения. Если читатель найдет решение этой задачи, его настоятельно просят опубликовать его; кроме того, автор данной книги будет очень признателен, если ему сообщат решение, как можно быстрее (при условии, что оно правильно).

Интерполируя по этой "логарифмической" шкале, можно прикинуть, что означает любая промежуточная оценка. Например, оценка 17 говорит о том, что данное упражнение чуть легче, чем упражнение средней трудности. Задача с оценкой 50, если она будет решена каким-либо читателем, в следующих изданиях данной книги может иметь уже оценку 45.

Автор честно старался давать объективные оценки, но тому, кто составляет задачи, трудно предвидеть, насколько трудными эти задачи окажутся для кого-то другого; к тому же у каждого человека существует определенный тип задач, которые он решает быстрее. Надеюсь, что выставленные мной оценки дают правильное представление о степени трудности задач, но в общем их нужно воспринимать как ориентировочные, а не абсолютные.

Эта книга написана для читателей самых разных степеней математической подготовки и искушенности, поэтому некоторые упражнения предназначены только для читателей с математическим уклоном. Если в каком-либо упражнении математические понятия или результаты используются более широко, чем это необходимо для тех, кого в первую очередь интересует программирование алгоритмов, то перед оценкой такого упражнения ставится буква "*M*". Если для решения упражнения требуется знание высшей математики в большем объеме, чем это дано в настоящей книге, то ставятся буквы "*BM*". Пометка "*BM*" отнюдь не является свидетельством того, что данное упражнение трудное.

Перед некоторыми упражнениями стоит стрелка ">"; это означает, что данное упражнение особенно поучительно и его рекомендуется обязательно выполнить. Само собой разумеется, никто не

ожидает, что читатель (или студент) будет решать все задачи, потому-то наиболее полезные из них и выделены. Это совсем не значит, что другие задачи не стоит решать! Каждый читатель должен по крайней мере попытаться решить все задачи с оценкой 10 и ниже; стрелки же помогут выбрать, какие задачи с более высокими оценками следует решить в первую очередь.

К большинству упражнений приведены ответы; они помещены в специальном разделе в конце книги. Пользуйтесь ими мудро; в ответ смотрите только после того, как вы приложили достаточно усилий, чтобы решить задачу самостоятельно, или же если для решения данной задачи у вас нет времени. Если получен собственный ответ, либо если вы действительно пытались решить задачу, только в этом случае ответ, помещенный в книге, будет поучительным и полезным. Как правило, ответы к задачам излагаются очень кратко, схематично, так как предполагается, что читатель уже честно пытался решить задачу собственными силами. Иногда в приведенном решении дается меньше информации, чем спрашивалось, чаще—наоборот. Вполне возможно, что полученный вами ответ окажется лучше ответа, помещенного в книге, или вы найдете ошибку в этом ответе; в таком случае автор был бы очень обязан, если бы вы как можно скорее подробно сообщили ему об этом. В последующих изданиях настоящей книги будет помещено уже исправленное решение вместе с именем его автора.

Сводка условных обозначений

$>$	Рекомендуется
M	С математическим уклоном
BM	Требует знания "высшей математики"
00	Требует немедленного ответа
10	Простое (на одну минуту)
20	Средней трудности (на четверть часа)
30	Повышенной трудности
40	Для "матпрактикума"
50	Исследовательская проблема

Упражнения

1. **[00]** Что означает пометка "*M20*"?
2. **[10]** Какое значение для читателя имеют упражнения, помещаемые в учебниках?
3. **[M50]** Докажите, что если n—целое число, $n > 2$, то уравнение $x^n + y^n = z^n$ неразрешимо в целых положительных числах x, y, z.

3. Случайные числа

Всякий, кто питает слабость к арифметическим методам получения случайных чисел, грешен вне всяких сомнений.
Джон фон Нейман (1951)

Круглые числа всегда фальшивы.
Сэмюэль Джонсон (около 1750)

Lest men suspect your tale untrue,
Keep probability in view.[1]
Джон Гэй (1727)

3.1. ВВЕДЕНИЕ

"Случайно выбранные" числа оказываются полезными для самых различных целей. Вот некоторые примеры:

a) *Моделирование.* Когда с помощью вычислительной машины моделируются природные явления, случайные числа позволяют приблизить модель к реальности. Моделирование применяется во многих областях: от ядерной физики (частицы испытывают случайные соударения) до системного анализа (скажем, люди входят в банк через случайные интервалы времени).

b) *Выборка.* Часто бывает, что проверка всех возможных вариантов практически неосуществима. Тогда на некоторые вопросы позволяет получить ответы случайная выборка.

[1] Чтобы люди поверили вашим россказням, помните о вероятности.—*Прим. перев.*

c) *Численный анализ.* Для решения сложных задач вычислительной математики была разработана остроумная техника, использующая случайные числа. Об этом написан ряд книг.

d) *Программирование для вычислительных машин.* Случайные значения служат хорошим источником данных при испытании эффективности различных алгоритмов для вычислительных машин. В этой книге нас будет в основном интересовать именно такое использование случайных чисел. Поэтому прежде чем пойдет речь о других алгоритмах, здесь, в третьей главе, будут рассмотрены случайные числа.

e) *Принятие решений.* Говорят, что многие руководители принимают решения, бросая монетку или кости. Ходят даже слухи, что некоторые профессора в колледжах добиваются успеха именно таким образом. Иногда бывает важно принимать совершенно непредвзятые решения. Полезно предусмотреть такую возможность для алгоритмов, применяемых в вычислительных машинах, например в случаях, когда принятие детерминированного решения может привести к замедлению счета. Случайность, кроме того,—существенная часть оптимальных стратегий в теории игр.

f) *Развлечения.* Многие проводят время, тасуя карты, бросая кости или наблюдая за колесом рулетки, и находят в этом неизъяснимое удовольствие. Таким традиционным использованием случайных чисел объясняется, почему термин "Монте-Карло" служит общим наименованием для всех алгоритмов, в которых применяют случайные числа.

Стало обычным в этом месте посвящать несколько абзацев философскому обсуждению того, что же такое "случайность". В некотором смысле такого объекта, как случайное число, просто нет. Скажем, двойка—это случайное число? Скорее мы говорим о *последовательности независимых* случайных чисел с определенным *законом распределения,* и это означает, грубо говоря, что каждое число было получено самым произвольным образом, без всякой связи с другими членами последовательности, и что у него есть определенная вероятность оказаться в любом заданном интервале.

Равномерным называется такое распределение, при котором каждое возможное число равновероятно. Обычно, если специально не оговорено что-либо иное, имеют в виду равномерные распределения.

Каждая из десяти цифр от 0 до 9 составляет примерно одну десятую часть всех цифр во всякой случайной (равномерной) последовательности цифр. Любая заданная пара двух соседних цифр должна составлять примерно $1/100$ часть всех пар, встречающихся в последовательности, и т. д. Тем не менее, если мы рассмотрим какую-нибудь конкретную случайную последовательность из миллиона цифр, в ней совсем не обязательно окажется ровно $100\,000$ нулей, $100\,000$ единиц и т. д. В действительности вероятность такого события очень мала. Закономерность же выполняется в среднем для *последовательности* таких последовательностей.

Любая заданная последовательность столь же вероятна, как и последовательность, состоящая из одних *нулей.* Более того, допустим, что мы выбираем случайным образом последовательность из миллиона цифр. Пусть оказалось, что первые $999\,999$ из них равны нулю. И в этом случае вероятность того, что последняя цифра будет нулем, все еще в точности равна $1/10$, если выборка действительно случайная. Для многих эти утверждения звучат как парадокс, но на самом деле в них нет противоречия.

Существует несколько способов сформулировать хорошее абстрактное определение случайной последовательности. Мы еще вернемся к этому интересному вопросу в § 3.5. Пока же достаточно интуитивно понять идею.

Раньше ученые, нуждавшиеся для своей работы в случайных числах, раскладывали карты, бросали кости или вытаскивали шары из урны, которую предварительно "как следует трясли". В 1927 г. Л. Типпетт опубликовал таблицы, содержащие свыше $40\,000$ случайных цифр, "произвольно взятых из отчетов о переписи". Позже были сконструированы специальные машины, механически вырабатывающие случайные числа. Первую такую машину в 1939 г. использовали М. Дж. Кендалл и Б. Бэбингтон-Смит при создании таблиц, включающих 100 тысяч случайных цифр. В 1955 г. компания RAND Corporation опубликовала хорошо известные таблицы с миллионом случайных цифр, полученных другой такой машиной. Известная машина ERNIE, вырабатывающая случайные числа, определяет выигравшие номера в Британской лотерее. [См. статьи Кендалла и Бэбингтон-Смита в *Journal of the Royal Statistical Society,* Series A, **101** (1938), 147–166, и Series B, **6** (1939), 51–61, а также обзор в *Math. Comp.,* **10** (1956), 39–43.]

Вскоре после создания вычислительных машин начались поиски эффективных методов получения случайных чисел, пригодных для использования в программах. В принципе можно работать и с таблицами, однако, этот метод имеет ограничения, связанные с конечным объемом памяти машин и затратами времени для ввода чисел в машину в том случае, когда таблица оказывается слишком короткой. Кроме того, довольно неприятно готовить таблицы заранее, да и вообще иметь с ними дело. Можно присоединить к ЭВМ машину типа ERNIE, но и этот путь оказывается неудовлетво-

рительным, потому что при отладке программы невозможно воспроизвести вторично вычисления, сделанные ранее.

Несовершенство всех этих методов пробудило интерес к получению случайных чисел с помощью арифметических операций вычислительной машины. Первым такой подход в 1946 г. предложил Джон фон Нейман, использовавший метод "середины квадрата". Идея заключается в том, что предыдущее случайное число возводится в квадрат, а затем из результата извлекаются средние цифры. Пусть, например, мы вырабатываем десятизначные числа и допустим, что предыдущее число было равно 5772156649; возведя его в квадрат, получим

$$33317792380594909201,$$

и поэтому следующее число равно 7923805949.

Метод вызывает довольно очевидное возражение. Как может быть случайной выработанная таким способом последовательность, если каждый ее член полностью определен своим предшественником?

Ответ заключается в том, что эта последовательность *не* случайна но *выглядит* как случайная. В типичных приложениях обычно не имеет значения, как связаны друг с другом два последующих числа последовательности; таким образом, неслучайный характер последовательности не является нежелательным. Интуитивно метод середины квадрата должен довольно хорошо "перемешивать" предыдущее число.

В научно-технической литературе последовательности, вырабатываемые детерминистскими способами, называются *псевдослучайными* или *квазислучайными*. Здесь же мы будем называть их просто случайными последовательностями, понимая, что они только *производят впечатление* случайных. Наверное, все, что можно сказать о случайной последовательности, это то, что она "по внешнему виду случайная". Точные математические определения случайности даются в § 3.5. Выработанные детерминистскими методами случайные числа оказались пригодными почти для всех приложений (хотя, конечно, они не могут заменить ERNIE в лотереях).

Однако первоначальный "метод середины квадрата" фон Неймана оказался сравнительно скудным источником случайных чисел. Недостаток его заключается в том, что последовательности имеют тенденцию превращаться в короткие циклы повторяющихся элементов. Например, если какой-нибудь член последовательности окажется равным нулю, все последующие члены также будут нулями. В начале пятидесятых годов некоторые ученые проводили эксперименты с методом середины квадрата. Дж. Э. Форсайт, работавший с четырехзначными (а не с десятизначными) числами, проверил 16 чисел в качестве начальных значений последовательностей Оказалось, что 12 из них порождали последовательности, оканчивающиеся циклом 6100, 2100, 4100, 8100, 6100, ..., а две последовательности выродились в нуль. Обширные эксперименты по исследованию метода середины квадрата провел Н. Метрополис, оперировавший главным образом двоичными числами. Работая с 20-разрядными числами, он показал, что существует тринадцать различных циклов, в которые могут выродиться последовательности; длина периода самого большого из них равна 142. Как только последовательность вырождается в нуль, довольно легко начать выработку случайных чисел заново. Гораздо трудней бороться с длинными циклами. Все же Р. Флойд (см. упр. 7) предложил остроумный метод, позволяющий зарегистрировать возникновение цикла в последовательности. Метод Флойда требует небольшой памяти машины, увеличивает время выработки случайного числа всего в три раза и сразу же сигнализирует, как только в последовательности появляется встречавшееся ранее число.

Теоретические недостатки метода середины квадрата обсуждаются в упр. 9 и 10. С другой стороны, отметим, что, работая с 38-разрядными двоичными числами, Н. Метрополис обнаружил последовательность, состоящую из 750 000 членов, отличающихся друг от друга. Статистические тесты подтвердили случайный характер полученной последовательности из 750000×38 битов. Это подтверждает, что, применяя метод середины квадрата, *можно* получить полезные результаты. Тем не менее без предварительных трудоемких вычислений ему не стоит излишне доверять.

Многие датчики случайных чисел, популярные сейчас, недостаточно хороши. Среди пользователей наметилась тенденция избегать их изучения. Довольно часто какой-нибудь старый сравнительно неудовлетворительный метод передается от одного программиста к другому вслепую, и сегодняшний пользователь уже ничего не знает об его недостатках. В этой главе мы убедимся, что нетрудно изучить самые важные свойства датчиков случайных чисел и научиться применять эти знания.

Изобрести простой датчик случайных чисел не так легко. Несколько лет назад этот факт произвел на автора большое впечатление. Он тогда пытался создать фантастически хороший датчик случайных чисел на основе следующего своеобразного метода.

Алгоритм К. *(Датчик "сверхслучайных" чисел.)* С помощью этого алгоритма данное десятизначное десятичное число X можно преобразовать в другое число, которое, как предполагается, является

следующим членом случайной последовательности. Казалось бы, алгоритм позволяет выработать достаточно случайную последовательность, но выяснилось, что это совсем не так. Причины неудачи разбираются ниже. (Читателю нет нужды слишком вникать в детали. Достаточно убедиться в большой сложности алгоритма.)

K1 [Выбрать число итераций.] Установить $Y \leftarrow \lfloor X/10^9 \rfloor$, задав его равным старшей цифре числа X. (Мы повторим $Y+1$ раз шаги с **K2** по **K13** включительно. Другими словами, случайное число будет вычисляться *случайное* число раз.)

K2 [Выбрать случайный шаг.] Установить $Z \leftarrow \lfloor X/10^8 \rfloor \bmod 10$, т. е. присвоить Z значение, равное второй по старшинству цифре числа X. Перейти к выполнению шага $\mathbf{K}(3+Z)$. (Другими словами, далее мы выполняем *случайно* выбранный шаг программы!)

K3 [Обеспечить $X \geq 5 \cdot 10^9$.] Если $X < 5 \cdot 10^9$, то установить $X \leftarrow X + 5 \cdot 10^9$.

K4 [Середина квадрата.] Заменить X числом $\lfloor X^2/10^5 \rfloor \bmod 10^{10}$, т. е. серединой квадрата числа X.

K5 [Умножить.] Заменить X на $(1001001001X) \bmod 10^{10}$.

K6 [Псевдодополнение.] Если $X < 10^8$, то установить $X \leftarrow X + 9814055677$, в противном случае $X \leftarrow 10^{10} - X$.

K7 [Переставить половинки.] Поменять местами 5 старших и 5 младших цифр, т. е. установить $X \leftarrow 10^5 \lfloor X \bmod 10^5 \rfloor + \lfloor X/10^5 \rfloor$, или, по-другому, взять средние 10 цифр числа $(10^{10}+1)X$.

K8 [Умножить.] (См. шаг **K5**.)

K9 [Уменьшить цифры.] Уменьшить на единицу каждую отличную от нуля цифру числа X (в десятичном представлении).

K10 [Модифицировать на 99999.] Если $X < 10^5$, то установить $X \leftarrow X^2 + 99999$, в противном случае $X \leftarrow X - 99999$.

K11 [Нормализовать.] (Здесь X не может быть равным нулю.) Если $X < 10^9$, то установить $X \leftarrow 10X$ и повторить этот шаг.

K12 [Модифицированная середина квадрата.] Заменить X на $\lfloor X(X-1)/10^5 \rfloor \bmod 10^{10}$, т. е. взять средние **10** цифр числа $X(X-1)$.

K13 [Повторить?] Если $Y > 0$, то уменьшить Y на 1 и вернуться на шаг **K2**. Если $Y = 0$, алгоритм завершен, причем текущее значение X считается искомым случайным числом. ∎

(Хотелось написать настолько сложную программу, реализующую описанный выше алгоритм, чтобы человек, читающий ее текст, не мог бы без объяснений догадаться, что же в ней делается.)

Учитывая все меры предосторожности, принятые в алгоритме K, не кажется ли вполне правдоподобным, что с его помощью можно получить бесконечное множество абсолютно случайных чисел? Нет! В действительности, как только этот алгоритм был реализован на ЭВМ, почти сразу же итерации сошлись к числу 6065038420, которое, в результате невероятного совпадения, преобразуется само в себя (см. табл. 1). При другом начальном значении последовательность, начиная с члена с номером 7401, повторяется с длиной периода 3178.

Мораль этой истории заключается в том, что *случайные числа нельзя вырабатывать с помощью случайно выбранного алгоритма.* Нужна какая-нибудь теория.

В этой главе мы рассмотрим методы выработки случайных чисел, превосходящие метод середины квадрата и алгоритм K в том отношении, что для них можно теоретически гарантировать выполнение определенных свойств случайной последовательности и отсутствие вырождения. Будут изложены некоторые детали, обеспечивающие такое случайное поведение, и уделено внимание технике применения случайных чисел. В частности, например, будет показано, как тасовать карты с помощью программы для ЭВМ.

Таблица 1

Колоссальное совпадение: число 6065038420 преобразуется в себя с помощью алгоритма К

Шаг	X (в конце шага)		Шаг	X (в конце шага)	
К1	6065038420		К9	1107855700	
К3	6065038420		К10	1107755701	
К4	6910360760		К11	1107755701	
К5	8031120760		К12	1226919902	$Y = 3$
К6	1968879240		К5	0048821902	
К7	7924019688		К6	9862877579	
К8	9631707688		К7	7757998628	
К9	8520606577		К8	2384626628	
К10	8520506578		К9	1273515517	
К11	8520506578		К10	1273415518	
К12	0323372207	$Y = 6$	К11	1273415518	
К6	9676627793		К12	5870802097	$Y = 2$
К7	2779396766		К11	5870802097	
К8	4942162766		К12	3172562687	$Y = 1$
К9	3831051655		К4	1540029446	
К10	3830951656		К5	7015475446	
К11	3830951656		К6	2984524554	
К12	1905867781	$Y = 5$	К7	2455429845	
К12	3319967479	$Y = 4$	К8	2730274845	
Кб	6680032521		К9	1620163734	
К7	3252166800		К10	1620063735	
К8	2218966800		К11	1620063735	
			К12	6065038420	$Y = 0$

Упражнения

>1. **[20]** Предположим, вы хотите, не используя ЭВМ, получить случайную десятичную цифру. Какой из перечисленных ниже методов вы предпочтете?

a) Откройте телефонный справочник в произвольном месте (т. е. ткните пальцем куда угодно) и возьмите младшую цифру первого попавшегося номера на выбранной вами странице.

b) Сделайте то же самое, но воспользуйтесь младшей цифрой номера *страницы*.

c) Бросьте кость в форме правильного икосаэдра, каждая из двадцати граней которого помечена цифрами $0, 0, 1, 1, \ldots, 9, 9$. Запишите цифру, которой будет помечена верхняя грань остановившейся кости. (Для эксперимента рекомендуется пользоваться столом с твердой обитой сукном поверхностью.)

d) Поставьте на минуту рядом с источником радиоактивного излучения счетчик Гейгера (примите меры предосторожности). Воспользуйтесь младшей цифрой числа отсчетов, показанного счетчиком. (Предполагается, что гейгеровский счетчик показывает число отсчетов в десятичном виде и перед началом эксперимента он был установлен на нуль.)

e) Взгляните на свои часы и, если секундная стрелка находится между числами $6n$ и $6(n + 1)$, выберите цифру n.

f) Попросите приятеля задумать случайную цифру и пусть он вам ее назовет.

g) Пусть то же самое сделает ваш недруг.

h) Предположим, что в забеге участвует 10 абсолютно неизвестных вам лошадей. Совершенно произвольно пронумеруйте их цифрами от 0 до 9. Воспользуйтесь номером победителя забега.

2. **[M22]** Какова вероятность обнаружить ровно $100\,000$ экземпляров любой заданной заранее цифры в случайной последовательности, состоящей из $1\,000\,000$ десятичных цифр?
3. **[10]** Какое число получится после применения метода середины квадрата к числу 1010101010?
4. **[10]** Почему при выполнении шага К11 алгоритма К значение X не может быть равным нулю? Что произошло бы с алгоритмом, если бы X могло быть нулем?
5. **[15]** Объясните, почему в любом случае нельзя ожидать получения "бесконечного множества" случайных чисел с помощью алгоритма К (даже если бы не произошло совпадения, приведенного в табл. 1), считая заранее известным тот факт, что любая последовательность, выработанная с использованием этого алгоритма, в конце концов станет периодической?

>**6.** **[M20]** Предположим, что мы хотим выработать последовательность целых чисел X_0, X_1, X_2, ... в интервале $0 \le X_n < m$. Пусть $f(x)$—любая функция, такая, что если $0 \le x < m$, то $0 \le f(x) < m$. Рассмотрим последовательность, полученную с помощью соотношения $X_{n+1} = f(X_n)$. (Примерами являются метод середины квадрата и алгоритм К.)

а) Покажите, что последовательность в конце концов становится периодической, т. е. существуют такие числа λ и μ, что значения X_0, X_1, ..., X_μ, ..., $X_{\mu+\lambda-1}$ различны, но $X_{n+\lambda} = X_n$, когда $n \ge \mu$. Найдите максимальное и минимальное возможные значения μ и λ.

b) Покажите, что существует $n > 0$, такое, что $X_n = X_{2n}$; наименьшее значение этого n лежит в интервале $\mu \le n \le \mu + \lambda$. Значение X_n является единственным в том смысле, что если $X_n = X_{2n}$ и $X_r = X_{2r}$, то $X_r = X_n$ (следовательно, $r - n$ кратно λ).

>**7.** **[20]** Примените результаты предыдущего упражнения, чтобы построить практичный алгоритм выработки случайных чисел, дополняющий датчик типа $X_{n+1} = f(X_n)$. Ваш алгоритм должен: а) обладать свойством приостанавливать выработку случайных чисел последовательности, как только повторяется ранее встречавшееся число, b) выработать по меньшей мере λ элементов перед остановкой, хотя значение λ заранее неизвестно, и с) использовать небольшую память (т. е. не разрешается. просто запоминать все вычисленные последовательные значения).

8. **[28]** Полностью проверьте метод середины квадрата для случая двузначных чисел. а) Мы можем начать с любого из 100 возможных значений 00, 01, ..., 99. В скольких случаях мы в конце концов придем к повторению цикла 00, 00, ...? [*Пример.* Начав с 43, мы получим последовательность 43, 84, 05, 02, 00, 00, 00, ...] b) Сколько может получиться различных циклов? Какова длина периода самого длинного цикла? с) Какое начальное значение позволяет получить самую длинную последовательность неповторяющихся элементов?

9. **[M14]** Докажите, что метод середины квадрата, использующий $2n$-значные числа с основанием b, имеет следующий недостаток: начиная с числа X, у которого старшие n цифр равны нулю, все последующие элементы последовательности становятся все меньше и меньше, пока не обратятся в нуль.

10. **[M16]** Сохранив предположения предыдущего упражнения, что можно сказать о последовательности элементов, следующих за числом X, у которого *самые младшие* n цифр равны нулю? Что, если младшие $(n + 1)$ цифр равны нулю?

>**11.** **[M26]** Рассмотрим случайные последовательности, вырабатываемые датчиками типа описанного в упр. 6. Если мы выбираем $f(x)$ и X_0 случайно и предполагаем, что любые из m^m возможных функций $f(x)$ и m начальных значений X_0 равновероятны, то какова вероятность того, что в конце концов последовательность выродится, образуя цикл с длиной периода $\lambda = 1$?

[*Замечание.* Предположения, принятые в этой задаче, дают возможность вполне естественным образом задуматься о "случайном" датчике случайных чисел подобного типа. Можно ожидать, что метод, подобный алгоритму К, должен быть похожим на средний датчик, рассмотренный здесь. Решение задачи дает меру того, насколько "колоссально" на самом деле совпадение, приведенное в табл. 1.]

>**12.** **[M31]** Используя предположения предыдущего упражнения, найдите среднюю длину цикла, который образуется в последовательностях. Какой средней длины достигает последовательность, прежде чем начать циклиться? (В обозначениях упр. 6 мы хотим определить средние значения λ и $\mu + \lambda$.)

13. **[M42]** Если $f(x)$ выбирается случайно в смысле упр. 11, какова средняя длина *самого длинного* цикла, полученного в результате изменения начального значения X_0? [*Замечание.* Ответ известен для специального случая, когда в качестве функции $f(x)$ рассматривается перестановка; см. упр. 1.3.3-23.]

14. **[M38]** Если $f(x)$ выбирается случайно в смысле упр. 11, каково среднее число различных циклов, получаемых в результате изменения начального значения? (Ср. с упр. 8(b).)

15. **[M15]** Если $f(x)$ выбирается случайно в смысле упр. 11, какова вероятность того, что ни один из циклов не имеет длины 1, безотносительно к выбору X_0?

16. **[15]** Последовательность, вырабатываемая с помощью метода, описанного в упр. 6, должна начать повторяться самое позднее после выработки m значений. Предположим, что мы обобщили метод так, что X_{n+1} теперь зависит не только от X_n, но и от X_{n-1}. Формально пусть $f(x, y)$ будет такая функция, что если $0 \le x, y < m$, то $0 \le f(x, y) < m$. Начинаем строить последовательность с произвольного выбора X_0 и X_1. Затем полагаем

$$X_{n+1} = f(X_n, X_{n-1}) \qquad \text{для } n > 0.$$

Какую максимальную длину периода можно получить в этом случае?

17. **[10]** Обобщите идею предыдущего упражнения так, чтобы X_{n+1} зависело от k предыдущих значений элементов последовательности.
18. **[M22]** Придумайте метод, аналогичный предложенному в упр. 7, для обнаружения цикла датчика случайных чисел общего вида, рассмотренного в предыдущем упражнении.
19. **[M50]** Решите задачи, поставленные в упр. 11–15, для более общего случая, когда X_{n+1} зависит от предыдущих k элементов последовательности. Каждая из m^{m^k} функций $f(x_1, x_2, \ldots, x_k)$ должна рассматриваться как равновероятная. [*Замечание.* Количество функций, дающих *максимальный* период, приводится в упр. 2.3.4.2-23.]

3.2. ВЫРАБОТКА РАВНОМЕРНО РАСПРЕДЕЛЕННЫХ СЛУЧАЙНЫХ ЧИСЕЛ

В этом параграфе мы рассмотрим методы получения последовательности случайных дробей, т. е. случайных *действительных чисел* U_n, *равномерно распределенных между нулем и единицей.* Так как в вычислительной машине действительное число всегда представляется с ограниченной точностью, фактически мы будем генерировать целые числа X_n в интервале от 0 до некоторого m. Тогда дробь

$$U_n = X_n/m \tag{1}$$

попадет в интервал от нуля до единицы. Обычно m на единицу больше максимального числа, которое можно записать в машинном слове [(m равно *размеру слова* (word size)]. Поэтому X_n можно интерпретировать (консервативно) как целое содержимое машинного слова с десятичной запятой, расположенной справа, а U_n можно считать (либерально) дробью, содержащейся в том же слове, с запятой в крайней левой позиции.

3.2.1. Линейный конгруэнтный метод

Наилучшие из известных сегодня датчиков случайных чисел представляют собой частные случаи следующей схемы, предложенной Д. Х. Лемером в 1948 г. [См. Proc. 2nd Symposium on Large-Scale Digital Computing Machinery (Cambridge: Harvard University Press, 1951), 141–146.] Выбираем четыре "магических числа":

$$\begin{array}{lll} X_0, & \text{начальное значение;} & X_0 \geq 0; \\ a, & \text{множитель;} & a \geq 0; \\ c, & \text{приращение;} & c \geq 0; \\ m, & \text{модуль;} & m > X_0, m > a, m > c. \end{array} \tag{1}$$

Тогда искомая последовательность случайных чисел $\langle X_n \rangle$ получается из соотношения

$$X_{n+1} = (aX_n + c) \bmod m, \qquad n \geq 0. \tag{2}$$

Она называется *линейной конгруэнтной последовательностью.*

Например, при $X_0 = a = c = 7$, $m = 10$ последовательность выглядит так:

$$7,\ 6,\ 9,\ 0,\ 7,\ 6,\ 9,\ 0,\ \ldots \tag{3}$$

Как видно из приведенного примера, последовательность не всегда оказывается "случайной", если выбирать X_0, a, c, m произвольно. В последующих разделах этой главы будут подробно исследованы принципы выбора этих значений.

Пример (3) иллюстрирует тот факт, что конгруэнтные последовательности всегда "зацикливаются", т. е. в конце концов числа образуют цикл, который повторяется бесконечное число раз. Это свойство присуще всем последовательностям, имеющим общий вид $X_{n+1} = f(X_n)$; см. упр. 3.1-6. Повторяющийся цикл называется *периодом.* Длина периода у последовательности (3) равна 4. Реальные последовательности, которыми пользуются, имеют, конечно, сравнительно большой период.

Специального упоминания заслуживает частный случай $c = 0$, когда процесс выработки случайных чисел происходит несколько быстрее. Позже мы увидим, что ограничение $c = 0$ уменьшает длину периода последовательности, но при этом все еще можно получить относительно большой период. В первоначальном методе Лемера было принято $c = 0$, хотя автор и упомянул возможность использования $c \neq 0$. Идея получения более длинных последовательностей за счет обобщения $c \neq 0$ принадлежит Томсону [*Comp. J.*, **1** (1958), 83, 86] и независимо Ротенбергу [*JACM*, **7** (1960), 75–77].

Многими авторами термины *мультипликативный конгруэнтный метод* и *смешанный конгруэнтный метод* применяются для обозначения линейных конгруэнтных методов с $c = 0$ и $c \neq 0$ соответственно.

Во всей этой главе буквы a, c, m, X_0 будут использоваться в том смысле, как это принято выше. Более того, чтобы упростить многие наши формулы, оказывается полезным определить

$$b = a - 1. \tag{4}$$

Можно сразу же отбросить случай $a = 1$, так как при этом $X_n = (X_0 + nc) \bmod m$, и очевидно, что последовательность не случайная. Вариант $a = 0$ еще хуже. Следовательно, для практических целей мы можем предположить, что

$$a \geq 2, \qquad b \geq 1. \tag{5}$$

Теперь можно обобщить соотношение (2),

$$X_{n+k} = (a^k X_n + (a^k - 1)c/b) \bmod m, \qquad k \geq 0, n \geq 0, \tag{6}$$

выразив $(n + k)$-й член прямо через n-й. (Следует обратить внимание на частный случай $n = 0$.) Последовательность, составленная из каждого k-го члена нашей последовательности образует другую линейную конгруэнтную последовательность с множителем a^k и приращением $((a^k - 1)c/b)$.

Упражнения

1. [10] В примере (3) иллюстрируется ситуация, когда $X_4 = X_0$, так что последовательность опять начинается сначала. Найдите пример такой линейной конгруэнтной последовательности с $m = 10$, в которой X_0 встречается только один раз.
>2. [M20] Покажите, что, если a и m взаимно простые, число X_0 будет периодически повторяться.
3. [M10] Объясните, почему, если a и m не являются взаимно простыми, последовательность в каком-то смысле неудачная и, вероятно, не слишком случайная. Поэтому, вообще говоря, мы будем стараться выбирать a и m взаимно простыми.
4. [11] Докажите соотношение (6).
5. [M20] Соотношение (6) выполняется для $k \geq 0$. Если это возможно, получите формулу, выражающую X_{n+k} через X_n и для *отрицательных* значений k.

3.2.1.1. Выбор модуля. Сначала обсудим, как правильно выбирать число m. Мы хотим, чтобы значение m было достаточно большим, так как длина периода не может быть больше m. (Даже если требуются только случайные нули и единицы, не следует брать $m = 2$, так как при этом в лучшем случае получится набор

$$\ldots, 0, 1, 0, 1, 0, 1, \ldots !$$

Методы использования равномерно распределенных случайных чисел для получения последовательности нулей и единиц обсуждаются в § 3.4.)

Другой фактор, влияющий на выбор m—это скорость выработки чисел: мы хотим подобрать такое значение, чтобы быстрее вычислять $(aX_n + c) \bmod m$.

Рассмотрим в качестве примера машину MIX. Мы можем вычислять $y \bmod m$, помещая y в регистры A и X, последующим делением на m. Предположив, что y и m—положительные числа, можно убедиться, что в результате в регистре X окажется $y \bmod m$. Но так как деление—сравнительно медленная операция, ее можно избежать за счет особо удобного выбора m, приняв его равным *размеру слова* (т. е. на единицу больше максимального целого числа, размещающегося в слове вычислительной машины).

Пусть w—такое максимальное целое число. Тогда операция сложения производится по модулю w, а умножение по модулю w также сравнительно просто, так как нужный результат получается в младших разрядах произведения. Таким образом, следующая программа эффективно вычисляет $(aX + c) \bmod w$:

```
01 LDA   A
02 MUL   X
03 SLAX  5
04 ADD   C
```
(1)

Результат получается в регистре A. В конце выполнения этой последовательности команд может произойти переполнение, сигнал которого можно выключить, написав вслед за последней командой "`JOV *+1`".

Менее широко известную тонкую технику можно использовать для вычислений по модулю $(w+1)$. По причинам, о которых будет сказано ниже, при $m = w + 1$ мы обычно полагаем $c = 0$. Поэтому нам

нужно вычислять просто $(aX) \bmod (w+1)$. Это делает следующая программа:

```
01 LDAN  X
02 MUL   A
03 STX   TEMP
04 SUB   TEMP
05 JANN  *+3
06 INCA  2
07 ADD   =W-1=
```
(2)

В регистре A теперь содержится значение $(aX) \bmod (w+1)$. Конечно, оно может быть любым в интервале между 0 и w. Поэтому читатель вполне законно может спросить, как можно записать так много значений с помощью A-регистра! (Очевидно, что в регистре не могут помещаться числа, большие, чем $w-1$.) Ответ состоит в том, что переполнение произойдет тогда и только тогда, когда результат равен w (в предположении, что сигнал переполнения был ранее выключен).

Для доказательства того, что программа (2), в самом деле, вычисляет $(aX) \bmod (w+1)$, заметим, что в строке 04 мы вычитаем младшую половину произведения из старшей. Переполнения при этом произойти не может, и, если $aX = qw + r$, где $0 \le r < w$, в регистре A после выполнения строки 04 окажется значение $r - q$. Заметим, что

$$aX = q(w+1) + (r-q),$$

а так как $q < w$, имеем $-w < r - q < w$. Следовательно, $(aX) \bmod (w+1)$ равняется либо $r-q$, либо $r-q+(w+1)$ в зависимости от того, какое из неравенств $r-q \ge 0$ или $r-q < 0$ выполняется. Если мы пользуемся конгруэнтным методом с $m = w+1$, отбрасывать значения w, если они получаются, бывает обычно легко, добавляя к программе (2) команды

```
01 JOV   *+1
00    ...
09 JNOV  *+3
10 LDAN  A
11 JMP   *-4
```
(3)

С помощью похожей техники можно вычислять произведение двух чисел по модулю $(w-1)$; см. упр. 8.

В последующих разделах нам потребуется уметь представлять m в виде произведения простых чисел. Это поможет правильно выбрать множитель a. В табл. 1 приводится факторизация чисел вида $w \pm 1$ почти для всех известных размеров слов. Более полные данные можно найти в книге Э. Каннингхэма и Х. Вудалла (Factorization of $y^n \pm 1$, Francis Hodgson, London, 1925) или получить с помощью методов п. 4.5.4.

Читатель может задать справедливый вопрос, почему нас так беспокоят значения $m = w \pm 1$, если удобства выбора $m = w$ настолько очевидны? Причина заключается в том, что при $m = w$ ***младшие цифры числа*** X_n ***намного менее случайны, чем старшие.*** Если d является делителем m и если

$$Y_n = X_n \bmod d \tag{4}$$

то, как можно легко показать,

$$Y_{n+1} = (aY_n + c) \bmod d \tag{5}$$

(ибо найдется некоторое целое q, для которого $X_{n+1} = aX_n + c - qm$, и поэтому если d является делителем m, то, переходя к сравнению по модулю d, получим (5)).[2]

Чтобы проиллюстрировать важность соотношения (5), предположим, например, что имеется двоичная вычислительная машина. Если $m = w = 2^e$, то младшие четыре бита X_n представляют число $Y_n = X_n \bmod 2^4$. Сущность формулы (5) заключается в том, что младшие четыре бита X_n образуют конгруэнтную последовательность с периодом, не превышающим 16. Точно так же младшие пять битов периодичны с периодом, равным самое большее 32. Самый же младший бит X_n либо сохраняет постоянное значение, либо строго периодически изменяется от нуля до единицы.

Подобной ситуации не возникает, когда $m = w \pm 1$. В этом случае младшие биты X_n ведут себя так же случайно, как и старшие. Например, члены последовательности при $w = 2^{35}$ и $m = 2^{35} - 1$

[2] Формула (4) определяет новую последовательность остатков $\gamma_0, \gamma_1, \ldots, \gamma_n, \ldots$, для которой выполнено (5).—*Прим. ред.*,

будут не слишком случайны, если вычислять вычеты по модулю 31, 71, 127 или 122921 (ср. с табл. 1). В то же время младшие биты будут вполне случайны (их можно рассматривать как вычеты членов последовательности, полученные при сравнении по модулю 2).

Таблица 1

Разложение $w \pm 1$ на простые множители

$2^e - 1$	e	$2^e + 1$
$7 \cdot 31 \cdot 151$	15	$3^2 \cdot 11 \cdot 331$
$3 \cdot 5 \cdot 17 \cdot 257$	16	65537
131071	17	$3 \cdot 43601$
$3^2 \cdot 7 \cdot 19 \cdot 73$	18	$5 \cdot 13 \cdot 37 \cdot 109$
524287	19	$3 \cdot 174763$
$3 \cdot 5^2 \cdot 11 \cdot 31 \cdot 41$	20	$17 \cdot 61681$
$7^2 \cdot 127 \cdot 337$	21	$3^2 \cdot 43 \cdot 5419$
$3 \cdot 23 \cdot 89 \cdot 683$	22	$5 \cdot 397 \cdot 2113$
$47 \cdot 178481$	23	$3 \cdot 2796203$
$3^2 \cdot 5 \cdot 7 \cdot 13 \cdot 17 \cdot 241$	24	$97 \cdot 257 \cdot 673$
$31 \cdot 601 \cdot 180l$	25	$3 \cdot 11 \cdot 251 \cdot 4051$
$3 \cdot 2731 \cdot 8191$	26	$5 \cdot 53 \cdot 157 \cdot 1613$
$7 \cdot 73 \cdot 262657$	27	$3^4 \cdot 19 \cdot 87211$
$3 \cdot 5 \cdot 29 \cdot 43 \cdot 113 \cdot 127$	28	$17 \cdot 15790321$
$233 \cdot 1103 \cdot 2089$	29	$3 \cdot 59 \cdot 3033169$
$3^2 \cdot 7 \cdot 11 \cdot 31 \cdot 151 \cdot 331$	30	$5^2 \cdot 13 \cdot 41 \cdot 61 \cdot 1321$
2147483647	31	$3 \cdot 715827883$
$3 \cdot 5 \cdot 17 \cdot 257 \cdot 65537$	32	$641 \cdot 6700417$
$7 \cdot 23 \cdot 89 \cdot 599479$	33	$3^2 \cdot 67 \cdot 683 \cdot 20857$
$3 \cdot 43691 \cdot 131071$	34	$5 \cdot 137 \cdot 953 \cdot 26317$
$31 \cdot 71 \cdot 127 \cdot 122921$	35	$3 \cdot 11 \cdot 43 \cdot 281 \cdot 86171$
$3^2 \cdot 5 \cdot 7 \cdot 13 \cdot 19 \cdot 37 \cdot 73 \cdot 109$	36	$17 \cdot 241 \cdot 433 \cdot 38737$
$223 \cdot 616318177$	37	$3 \cdot 1777 \cdot 25781083$
$3 \cdot 174763 \cdot 524287$	38	$5 \cdot 229 \cdot 457 \cdot 525313$
$7 \cdot 79 \cdot 8191 \cdot 121369$	39	$3^2 \cdot 2731 \cdot 22366891$
$3 \cdot 5^2 \cdot 11 \cdot 17 \cdot 31 \cdot 41 \cdot 61681$	40	$257 \cdot 4278255361$
$13367 \cdot 164511353$	41	$3 \cdot 83 \cdot 8831418697$
$3^2 \cdot 7^2 \cdot 43 \cdot 127 \cdot 337 \cdot 5419$	42	$5 \cdot 13 \cdot 29 \cdot 113 \cdot 1429 \cdot 14449$
$431 \cdot 9719 \cdot 2099863$	43	$3 \cdot 2932031007403$
$3 \cdot 5 \cdot 23 \cdot 89 \cdot 397 \cdot 683 \cdot 2113$	44	$17 \cdot 353 \cdot 2931542417$
$7 \cdot 31 \cdot 73 \cdot 151 \cdot 631 \cdot 23311$	45	$3^3 \cdot 11 \cdot 19 \cdot 331 \cdot 18837001$
$3 \cdot 47 \cdot 178481 \cdot 2796203$	46	$5 \cdot 277 \cdot 1013 \cdot 1657 \cdot 30269$
$2351 \cdot 4513 \cdot 13264529$	47	$3 \cdot 283 \cdot 165768537521$
$3^2 \cdot 5 \cdot 7 \cdot 13 \cdot 17 \cdot 97 \cdot 241 \cdot 257 \cdot 673$	48	$193 \cdot 65537 \cdot 22253377$
$179951 \cdot 3203431780337$	59	$3 \cdot 2833 \cdot 37171 \cdot 1824726041$
$3^2 \cdot 5^2 \cdot 7 \cdot 11 \cdot 13 \cdot 31 \cdot 41 \cdot 61 \cdot 151 \cdot 331 \cdot 1321$	60	$17 \cdot 241 \cdot 61681 \cdot 4562284561$
$7^2 \cdot 73 \cdot 127 \cdot 337 \cdot 92737 \cdot 649657$	63	$3^3 \cdot 19 \cdot 43 \cdot 5419 \cdot 77158673929$
$3 \cdot 5 \cdot 17 \cdot 257 \cdot 641 \cdot 65537 \cdot 6700417$	64	$274177 \cdot 67280421310721$

$10^e - 1$	e	$10^e + 1$
$3^3 \cdot 7 \cdot 11 \cdot 13 \cdot 37$	6	$101 \cdot 9901$
$3^2 \cdot 239 \cdot 4649$	7	$11 \cdot 909091$
$3^2 \cdot 11 \cdot 73 \cdot 101 \cdot 137$	8	$17 \cdot 5882353$
$3^4 \cdot 37 \cdot 333667$	9	$7 \cdot 11 \cdot 13 \cdot 19 \cdot 52579$
$3^2 \cdot 11 \cdot 41 \cdot 271 \cdot 9091$	10	$101 \cdot 3541 \cdot 27961$
$3^2 \cdot 21649 \cdot 513239$	11	$11^2 \cdot 23 \cdot 4093 \cdot 8779$
$3^3 \cdot 7 \cdot 11 \cdot 13 \cdot 37 \cdot 101 \cdot 9901$	12	$73 \cdot 137 \cdot 99990001$
$3^2 \cdot 11 \cdot 17 \cdot 73 \cdot 101 \cdot 137 \cdot 5882353$	16	$353 \cdot 449 \cdot 641 \cdot 1409 \cdot 69857$

Другая возможность—это выбор в качестве m наибольшего простого числа, меньшего, чем w. Его можно найти, применяя технику п. 4.5.4, в котором есть и таблица соответствующих больших простых чисел.

Для большинства приложений младшие биты не существенны, и выбор $m = w$ является вполне удовлетворительным при условии, что программист, пользующийся случайными числами, делает это разумно.

Упражнения

1. **[12]** Из упр. 3.2.1-3 мы заключили, что у самых лучших конгруэнтных датчиков числа a и m взаимно простые. Покажите, что в этом случае можно получить $(aX + c) \bmod w$ в регистре X с помощью всего *трех* команд MIX, а не четырех, как в примере (1).
2. **[16]** Напишите подпрограмму MIX, удовлетворяющую трем условиям:

   Обращение: `JMP RANDM`
   Условия входа: В ячейке `XRAND` содержится целое число X.
   Условия выхода: $X \leftarrow$ `rA` $\leftarrow (aX + c) \bmod w$, `rX` $\leftarrow 0$, переполнение "выключено".

   (Таким образом, в результате обращения к этой подпрограмме будет получено следующее случайное число линейной конгруэнтной последовательности.)
3. **[20]** Как описать на языке автокода MIX константу $w - 1$ независимо от размера байта?
4. **[10]** Объясните смысл последовательности команд (3).
5. **[20]** Дано, что m меньше максимального целого, помещающегося в слове, а x и y — неотрицательные целые числа, меньшие, чем m. Покажите, что для вычисления разности $(x - y) \bmod m$ достаточно четырех команд MIX и не нужна операция деления. Какова наилучшая программа для вычисления суммы $(x + y) \bmod m$?

>6. **[20]** В предыдущем упражнении предполагается, что вычитание по $\bmod m$ проще, чем сложение по $\bmod m$. Рассмотрите последовательности, образующиеся по правилу

$$X_{n+1} = (aX_n - c) \bmod m.$$

Существенно ли они отличаются от линейных конгруэнтных последовательностей, определенных в тексте? Удобнее ли они для эффективного вычисления на машинах?

7. **[M24]** Какие характерные особенности табл. 1 вы можете отметить?

>8. **[20]** Напишите MIX-программу для вычисления $(aX) \bmod (w - 1)$, аналогичную (2). Значения 0 и $w - 1$ на входе и выходе вашей программы следует считать эквивалентными.
9. **[23]** Напишите MIX-программу для вычисления $(aX) \bmod (w - 2)$, аналогичную требующейся в упр. 8.

3.2.1.2. Выбор множителя. В этом подпункте мы покажем, как выбирать множитель a, чтобы получать *период максимальной длины*. Для любой последовательности, предназначенной для использования в качестве источника случайных чисел, важен большой период. Действительно, нам бы хотелось, чтобы период содержал значительно больше чисел, чем это необходимо для решения какой-либо одной задачи. Поэтому займемся здесь длиной периода. Читателю следует, однако, помнить, что большой период—это только один из необходимых признаков случайности нашей последовательности. Вполне возможны абсолютно неслучайные последовательности с очень большим периодом. Например, при $a = c = 1$ последовательность сводится просто к $X_{n+1} = (X_n + 1) \bmod m$. Очевидно, ее период равен m, тем не менее ее никак нельзя назвать случайной. Другие соображения, влияющие на выбор множителя, будут приведены в этой главе позже.

Так как возможны только m различных значений, длина периода не может превышать m. Достижима ли максимальная длина m? Приведенный выше пример показывает, что да, хотя выбор $a = c = 1$ и не приводит к желаемой последовательности. Исследуем *все* возможные способы выбора a и c, которые дают период длины m. [*Замечание.* Когда период имеет длину m, каждое число от 0 до $(m - 1)$ встречается за период ровно один раз. Поэтому в этом случае выбор X_0 не влияет на длину периода.]

Теорема A. *Длина периода линейной конгруэнтной последовательности равна m тогда и только тогда, когда*

i) *c и m—взаимно простые числа;*
ii) *$b = a - 1$ кратно p для любого простого p, являющегося делителем m;*
iii) *b кратно 4, если m кратно 4.*

Идеи доказательства известны не менее ста лет. Первое доказательство теоремы в данной формулировке было дано М. Гринбергером для частного случая $m = 2^e$ (*JACM*, **8** (1961), 383–389). Достаточность условий (i), (ii), (iii) в общем случае была показана Халлом и Добеллом (*SIAM Review*, **4** (1962), 230–254). Для доказательства теоремы рассмотрим сначала некоторые вспомогательные результаты теории чисел, которые интересны и сами по себе.

Лемма P. *Пусть p—простое число, а e—положительное целое, такое, что $p^e > 2$. Если*

$$x \equiv 1 \pmod{p^e}, \quad x \not\equiv 1 \pmod{p^{e+1}}, \tag{1}$$

то

$$x^p \equiv 1 \pmod{p^{e+1}}, \quad x^p \not\equiv 1 \pmod{p^{e+2}}. \tag{2}$$

Доказательство. Мы имеем $x = 1 + qp^e$, где q—некоторое целое, не кратное p. По формуле бинома запишем

$$\begin{aligned} x^p &= 1 + \binom{p}{1}qp^e + \cdots + \binom{p}{p-1}q^{p-1}p^{(p-1)e} + q^p p^{pe} = \\ &= 1 + qp^{e+1}\left(1 + \frac{1}{p}\binom{p}{2}qp^e + \frac{1}{p}\binom{p}{3}q^2p^{2e} + \cdots + \frac{1}{p}\binom{p}{p}q^{p-1}p^{(p-1)e}\right). \end{aligned}$$

Выражение в скобках—целое число, причем каждое слагаемое кратно p, за исключением первого члена. В самом деле, если $1 < k < p$, то $\binom{p}{k}$ делится на p (ср. с упр. 1.2.6-10). Следовательно,

$$\frac{1}{p}\binom{p}{k}q^{k-1}p^{(k-1)e}$$

делится на $p^{(k-1)e}$. Последний член $q^{p-1}p^{(p-1)e-1}$ делится на p, так как $(p-1)e > 1$ при $p^e > 2$. Таким образом, $x^p = 1 + q'p^{e+1}$, где q'—целое число, которое не делится на p. Тем самым доказательство завершено. (*Замечание:* обобщение этого результата содержится в упр. 3.2.2-11 (a).) ∎

Лемма Q. *Пусть разложение m на простые множители имеет вид*

$$m = p_1^{e_1} \dots p_t^{e_t}. \tag{3}$$

Длина λ периода линейной конгруэнтной последовательности, определяемой (X_0, a, c, m), равна наименьшему общему кратному длин λ_j периодов линейных конгруэнтных последовательностей

$$(X_0 \bmod p_j^{e_j}, ap_j^{e_j}, cp_j^{e_j}, p_j^{e_j}), \qquad 1 \le j \le t.$$

Доказательство. Индукцией по t достаточно доказать, что, если r и s—взаимно простые числа, длина λ периода линейной конгруэнтной последовательности, определяемой (X_0, a, c, rs), равна наименьшему общему кратному длин λ_1 и λ_2 периодов последовательностей $(X_0 \bmod r, a \bmod r, c \bmod r, r)$ и $(X_0 \bmod s, a \bmod s, c \bmod s, s)$.

В предыдущем разделе (см. формулу (5)) мы видели, что если обозначить элементы этих трех последовательностей соответственно X_n, Y_n, Z_n, то выполняются следующие соотношения:

$$Y_n = X_n \bmod r \text{ и } Z_n = X_n \bmod s \qquad \text{для всех } n \ge 0.$$

Поэтому из свойства D п. 1.2.4 находим, что

$$X_n = X_k \qquad \text{тогда и только тогда, когда } Y_n = Y_k \text{ и } Z_n = Z_k. \tag{4}$$

Пусть λ'—наименьшее общее кратное длин λ_1 и λ_2; докажем, что $\lambda' = \lambda$. Так как $X_n = X_{n+\lambda}$ для всех достаточно больших n, имеем $Y_n = Y_{n+\lambda}$ (следовательно, λ кратно λ_1) и $Z_n = Z_{n+\lambda}$ (следовательно, λ кратно λ_2), и поэтому $\lambda \ge \lambda'$. С другой стороны, мы знаем, что $Y_n = Y_{n+\lambda'}$ и $Z_n = Z_{n+\lambda'}$ для всех достаточно больших n. Поэтому на основании (4) имеем $X_n = X_{n+\lambda'}$, из чего получаем $\lambda \le \lambda'$, так что $\lambda = \lambda'$. ∎

Теперь мы готовы доказать теорему A. Имея в виду лемму Q, достаточно доказать теорему для случая, когда m есть степень простого числа. В самом деле, неравенство

$$p_1^{e_1} \dots p_t^{e_t} = \lambda = \text{нок}(\lambda_1, \dots, \lambda_t) \le \lambda_1 \dots \lambda_t \le p_1^{e_1} \dots p_t^{e_t}$$

может быть справедливо в том и только том случае, если $\lambda_j = p_j^{e_j}$ для $1 \le j \le t$.

Поэтому предположим, что $m = p^e$, где p—простое, а e—положительное целое число. Очевидно, что при $a = 1$ теорема справедлива, поэтому можно считать $a > 1$. Длина периода равна m тогда и

только тогда, когда каждое целое число x, такое, что $0 \le x < m$, встречается в последовательности на протяжении периода (поскольку ни одно из значений не может встретиться за период более одного раза). Таким образом, период имеет длину m в том и только том случае, если длина периода последовательности с $X_0 = 0$ равна m, что оправдывает предположение $X_0 = 0$. Из. формулы 3.2.1-(6) имеем

$$X_n = \left(\frac{a^n - 1}{a - 1}\right) c \bmod m. \tag{5}$$

Если числа c и m не являются взаимно простыми, X_n никогда не может быть равно 1. Поэтому условие (i) оказывается необходимым. Период имеет длину m тогда и только тогда, когда наименьшее положительное значение n, для которого $X_n = X_0 = 0$, таково, что $n = m$. Теперь в силу (5) и условия (i) наша теорема сводится к доказательству следующего утверждения.

Лемма R. *Предположим, что $1 < a < p^e$, где p—простое число. Если λ—наименьшее положительное целое, для которого $(a^\lambda - 1)/(a - 1) \equiv 0 \pmod{p^e}$, то*

$$\lambda = p^e \text{ тогда и только тогда, когда } \begin{cases} a \equiv 1 \pmod{p} & \text{при } p > 2, \\ a \equiv 1 \pmod{4} & \text{при } p = 2. \end{cases}$$

Доказательство. Предположим, что $\lambda = p^e$. Если $a \not\equiv 1 \pmod{p}$, то $(a^n - 1)/(a - 1) \equiv 0 \pmod{p^e}$ в том и только том случае, когда $a^n - 1 \equiv 0 \pmod{p^e}$. Условие $a^{p^e} - 1 \equiv 0 \pmod{p^e}$ тогда требует, чтобы выполнялось соотношение $a^{p^e} \equiv 1 \pmod{p}$. Но из теоремы 1.2.4F следует, что $a^{p^e} \equiv a \pmod{p}$; таким образом получаем $a \not\equiv 1 \pmod{p}$, что приводит к противоречию. Если же $p = 2$ и $a \equiv 3 \pmod{4}$, то из упр. 8 имеем $(a^{2^{e-1}} - 1)/(a - 1) \equiv 0 \pmod{2^e}$. Эти соображения обосновывают необходимость равенства $a = 1 + qp^f$, где $p^f > 2$, а q не кратно p для любых $\lambda = p^e$.

Остается показать, что это условие *достаточно* для выполнения соотношения $\lambda = p^e$. Повторно используя лемму P, находим, что

$$a^{p^g} \equiv 1 \pmod{p^{f+g}}, \quad a^{p^g} \not\equiv 1 \pmod{p^{f+g+1}},$$

и поэтому

$$\begin{aligned} (a^{p^g} - 1)/(a - 1) &\equiv 0 \pmod{p^g}, \\ (a^{p^g} - 1)/(a - 1) &\not\equiv 0 \pmod{p^{g+1}}. \end{aligned} \tag{6}$$

В частности, $(a^{p^e} - 1)/(a - 1) \equiv 0 \pmod{p^e}$. Таким образом, в конгруэнтной последовательности $(0, a, 1, p^e)$ имеем $X_n = (a^n - 1)/(a - 1) \bmod p^e$; поэтому длина ее периода равна λ, т. е. $X_n = 0$ тогда и только тогда, когда n кратно λ. Следовательно, p^e кратно λ. Это возможно только, если $\lambda = p^g$ для некоторого g. Из соотношения (6) получаем $\lambda = p^e$, что завершает доказательство. ∎

Теперь закончено и доказательство теоремы A. ∎

В заключение этого раздела рассмотрим специальный случай чисто мультипликативных датчиков, для которых $c = 0$. Хотя при этом выработка случайных чисел происходит несколько быстрее, теорема A показывает, что добиться максимальной длины периода нельзя. Действительно, это вполне очевидно, так как члены последовательности удовлетворяют соотношению

$$X_{n+1} = aX_n \bmod m, \tag{7}$$

и значение $X_n = 0$ может в ней встретиться, только если последовательность вырождается в нуль. Вообще, если d—любой делитель m и если X_n кратно d, все последующие значения X_{n+1}, X_{n+2}, ... тоже кратны d. Поэтому при $c = 0$ желательно, чтобы X_n были взаимно просты с m для всех n, а это ограничивает длину периода.

Можно добиться приемлемо большого периода, даже если мы настаиваем, чтобы $c = 0$. Попытаемся теперь найти такие условия, определяющие множитель, чтобы и в этом частном случае длина периода была максимальна.

Вследствие леммы Q период последовательности полностью определяется периодами последовательностей с $m = p^e$, так что рассмотрим именно такую ситуацию. Мы имеем $X_n = a^n X_0 \bmod p^e$, и ясно, что длина периода равна 1, если a кратно p^3. Поэтому выберем a взаимно простым с p^e. Тогда период равен наименьшему целому λ, такому, что $X_0 = a^\lambda X_0 \bmod p^e$. Если наибольший общий делитель X_0 и p^e есть p^f, это условие эквивалентно следующему:

$$a^\lambda \equiv 1 \pmod{p^{e-f}}. \tag{8}$$

[3] Если a кратно p, то, вообще говоря, период не превосходит e.—*Прим. ред.*

По теореме Эйлера (упр. 1.2.4-28)

$$a^{\varphi(p^{e-f})} \equiv 1 \pmod{p^{e-f}};$$

следовательно, λ есть делитель:

$$\varphi(p^{e-f}) = p^{e-f-1}(p-1).$$

Для взаимно простых a и m наименьшее целое λ, для которого $a^\lambda \equiv 1 \pmod{m}$, обычно называют *порядком по модулю* m. Величина a, которой соответствует *максимально* возможный порядок по модулю m, называется *первообразным элементом* по модулю m[4].

Обозначим через $\lambda(m)$ порядок первообразного элемента, т. е. максимально возможный порядок по модулю m. Предыдущие замечания показывают, что $\lambda(p^e)$ есть делитель $p^{e-1}(p-1)$; не составляет труда (см. ниже упр. 11–16) привести точные значения $\lambda(m)$ в следующих случаях:

$$\begin{aligned}
&\lambda(2) = 1, \quad \lambda(4) = 2, \quad \lambda(2^e) = 2^{e-2}, && \text{если } e \ge 3,\\
&\lambda(p^e) = p^{e-1}(p-1), && \text{если } p > 2,\\
&\lambda(p_1^{e_1} \ldots p_t^{e_t}) = \text{нок}(\lambda(p_1^{e_1}, \ldots, \lambda(p_t^{e_t}))).
\end{aligned} \tag{9}$$

Наши замечания можно суммировать в следующей теореме:

Теорема B. *(P. Кармайкл) [R. D. Carmichael, Bull. Amer. Math. Soc.,* **16** *(1910), 232–238]. Максимально возможный при $c = 0$ период равен $\lambda(m)$, где $\lambda(m)$ определяется выражениями (9). Такой период реализуется, если*

i) X_0 *и* m*—взаимно простые числа;*
ii) a*—первообразный элемент по модулю* m*.* ∎

Заметьте, что если m—простое число, можно получить длину периода $m-1$, что всего лишь на единицу меньше максимального значения.

Вопрос теперь заключается в том, как находить первообразные элементы по модулю m. Из упражнений в конце параграфа вытекает

Теорема C. *Число a есть первообразный элемент по модулю p^e тогда и только тогда, когда*

i) $p^e = 2$, a*—нечетное; или* $p^e = 4$, $a \bmod 4 = 3$*; или* $p^e = 8$, $a \bmod 8 = 3, 5, 7$*; или* $p = 2$, $e \ge 4$, $a \bmod 8 = 3$ *или* 5*;*

или

ii) p*—нечетное,* $e = 1$, $a \not\equiv 0 \pmod{p}$ *и* $a^{(p-1)q} \not\equiv 1 \pmod{p}$ *для любого простого делителя* q *числа* $p-1$*;*

или

iii) p*—нечетное,* $e > 1$, a *удовлетворяет (ii) и* $a^{p-1} \not\equiv 1 \pmod{p^2}$[5]*.* ∎

Условия (ii) и (iii) этой теоремы легко проверяются на вычислительной машине для больших значений p, если для вычисления степеней использовать эффективные методы, обсуждаемые в п. 4.6.3. Наконец, если нам даны величины a_j, являющиеся первообразными элементами по модулю $p_j^{e_j}$, можно найти единственное значение a, такое, что $a \equiv a_j \pmod{p_j^{e_j}}$, $1 \le j \le t$, применяя "китайскую теорему об остатках", которая обсуждается в п.4.3.2. Следовательно, a будет первообразным элементом по модулю $p_1^{e_1} \ldots p_t^{e_t}$. Это дает нам довольно эффективный способ вычисления множителей, удовлетворяющих условию теоремы B, для любого значения m. Однако в общем случае вычисления оказываются несколько длинными.

Для важного случая $m = 2^e$ при $e \ge 4$ приведенные выше условия сводятся к единственному простому требованию, чтобы $a \equiv 3$ или $5 \pmod{8}$. В этом случае четвертая часть всех возможных множителей дает максимальный период.

Второй наиболее распространенный случай, это $m = 10^e$. Пользуясь леммами P и Q, нетрудно получить необходимые и достаточные условия достижения максимального периода для десятичной вычислительной машины.

---

[4] Не следует путать первообразный элемент с первообразным корнем. Первообразные корни существуют не для всех m.—*Прим. ред.*

[5] *Подразумевается, что p— простое; если модуль имеет вид $2, 2^2, p^e$, где p—нечетное, первообразные элементы будут и первообразными корнями.—Прим. ред.*

Теорема D. *Если* $m = 10^e$, $e \geq 5$, $c = 0$ *и* X_0 *не кратно* 2 *или* 5, *период линейной конгруэнтной последовательности равен* $5 \times 10^{e-2}$ *в том и только том случае, когда* $a \bmod 200$ *принимает одно из следующих* 32 *значений:*

$$3, 11, 13, 19, 21, 27, 29, 37, 53, 59, 61, 67, 69, 77, 83, 91, 109, 117,$$
$$123, 131, 133, 139, 141, 147, 163, 171, 173, 179, 181, 187, 189.197. \blacksquare \tag{10}$$

Упражнения

1. **[10]** Какова длина периода линейной конгруэнтной последовательности с $X_0 = 5772156648$, $a = 3141592621$, $c = 2718281829$, $m = 10000000000$?
2. **[10]** Гарантирует ли выполнение следующих двух условий: (i) c нечетное, (ii) $a \bmod 4 = 1$, максимальную длину периода, когда $m = 2^e$, т. е. степень двойки?
3. **[13]** Предположим, что $m = 10^e$, где $e \geq 3$, а c не кратно ни 2, ни 5. Покажите, что линейная конгруэнтная последовательность будет иметь максимально большой период тогда и только тогда, когда $a \bmod 20 = 1$.
4. **[20]** Чему равно значение $X_{2^{e-1}}$, если a и c удовлетворяют условиям теоремы A, $m = 2^e$, $X_0 = 0$?

>5. **[20]** Найдите все множители a, удовлетворяющие условиям теоремы A, когда $m = 2^{35} + 1$ (простые множители числа m можно найти из табл. 3.2.1.1-1).

6. **[20]** Найдите все множители a, удовлетворяющие условиям теоремы A, при $m = 10^6 - 1$ (см. табл. 3.2.1.1-1).

>7. **[M24]** Период конгруэнтной последовательности не обязательно начинать с X_0. Однако мы всегда можем найти индексы $\mu \geq 0$, $\lambda > 0$, такие, что $X_{n+\lambda} = X_n$ для любых $n \geq \mu$, причем μ и λ—наименьшие значения, обладающие этим свойством. (Ср. с упр. 3.1-6 и 3.2.1-1.) Пусть μ_j, λ_j соответствуют последовательности $(X_0 \bmod p_j^{e_j}, a \bmod p_j^{e_j}, c \bmod p_j^{e_j}, p_j^{e_j})$ и μ, λ соответствуют последовательности $(X_0, a, c, p_1^{e_1} \ldots p_t^{e_t})$; лемма Q устанавливает, что λ есть наименьшее общее кратное для $\lambda_1, \ldots, \lambda_t$. Чему равно значение μ, выраженное через $\mu_1, \ldots, \mu_t$? Какое максимальное значение μ можно получить, изменяя X_0, a и c при фиксированном $m = p_1^{e_1} \ldots p_t^{e_t}$?

8. **[M20]** Покажите, что если $a \bmod 4 = 3$, $e > 1$, то $(a^{2^{e-1}} - 1)/(a - 1) \equiv 0 \pmod{2^e}$. (Используйте лемму P.)

>9. **[M22] (У. Томсон.)** Для $c = 0$ и $m = 2^e \geq 8$ теоремы B и C утверждают, что длина периода равна 2^{e-2} в том и только том случае, когда множитель a удовлетворяет соотношениям $a \bmod 8 = 3$ или $a \bmod 8 = 5$. Покажите, что каждая такая последовательность в сущности является линейной конгруэнтной последовательностью с $m = 2^{e-2}$, имеющей *полный* период в следующем смысле:

a) если $X_{n+1} = (4c + 1)X_n \bmod 2^e$ и $X_n = 4Y_n + 1$, то

$$Y_{n+1} = ((4c + 1)Y_n + c) \bmod 2^{e-2};$$

b) если $X_{n+1} = (4c - 1)X_n \bmod 2^e$ и $X_n = ((-1)^n(4Y_n + 1)) \bmod 2^e$, то

$$Y_{n+1} = ((1 - 4c)Y_n - c) \bmod 2^{e-2}.$$

[*Замечание.* В этих формулах c—нечетное целое. В литературе можно встретить утверждения о том, что последовательности с $c = 0$, удовлетворяющие условиям теоремы B, несколько более случайны, чем те, которые удовлетворяют условиям теоремы A, несмотря на то, что в случае теоремы B период в четыре раза меньше. Это упражнение опровергает подобные утверждения.]

10. **[M21]** Для каких значений m справедливо равенство $\lambda(m) = \varphi(m)$?
11. **[M28]** Пусть x—нечетное целое число, большее 1.

(a) Покажите, что существует единственное целое $f > 1$, такое, что $x \equiv 2^f \pm 1 \pmod{2^{f+1}}$.

(b) При условии, что $1 < x < 2^e - 1$, а f—соответствующее целое число из п.(a), покажите, что порядок $x \pmod{2^e}$ равен 2^{e-f}.

(c) В частности, это доказывает теорему C(i).

12. **[M26]** Пусть p—нечетное простое число Докажите что если $e > 1$, то a—первообразный элемент по модулю p^e в том и только том случае, когда a—первообразный элемент по модулю p и $a^{p-1} \equiv 1 \pmod{p^2}$. (Предположите, что $\lambda(p^e) = p^{e-1}(p - 1)$. Это доказывается ниже в упр. 14 и 16.
13. **[M22]** Пусть p—простое число и a не является первообразным элементом по модулю p. Покажите, что либо a кратно p либо $a^{(p-1)/q} \equiv 1 \pmod{p}$ для некоторого простого числа q, делящего $p - 1$.
14. **[M18]** Пусть $e > 1$, p—нечетное простое и a—первообразный элемент по модулю p, докажите, что тогда либо a, либо $a + p$—первообразный элемент по модулю p^e. [*Указание:* см. упр. 12.]

15. **[M29]** (a) Пусть a_1, a_2 взаимно простые с m. Пусть для этих чисел порядки по модулю m равны λ_1, λ_2 соответственно. Докажите что если λ—наименьшее общее кратное λ_1 и λ_2, то $a_1^{\kappa_1}a_2^{\kappa_2}$ имеет порядок λ по модулю m для выбранных надлежащим образом κ_1, κ_2. [*Указание.* Рассмотрите сначала случай, когда λ_1 и λ_2—взаимно простые числа.] (b) Пусть $\lambda(m)$—максимальный по всем элементам порядок по модулю m. Докажите, что $\lambda(m)$ кратно порядку по модулю m для любого элемента. Другими словами, докажите, что $a^{\lambda(m)} \equiv 1 \pmod m$ для любого a, взаимно простого с m.

>16. **[M24]** Пусть p—простое число. (a) Пусть $f(x) = x^n + c_1x^{n-1} + \cdots + c_n$, где все c—целые числа, и задано такое целое a, что $f(a) \equiv 0 \pmod p$. Покажите, что существует полином $q(x) = x^n + q_1x^{n-2} + \cdots + q_{n-1}$ с целыми коэффициентами, такой, что $f(x) \equiv (x-a)q(x) \pmod p$ для всех целых x. (b) Пусть $f(x)$—полином, такой же, как и в (a). Покажите, что $f(x)$ имеет самое большее n различных "корней" по модулю p, т.е. существует самое большее n целых чисел a, таких, что $f(a) \equiv 0 \pmod p$, $0 \le a < p$. (c) В упр. 15(b) утверждается, что полином $f(x) = x^{\lambda(p)} - 1$ имеет $p-1$ различных корней; следовательно, существует целое a с порядком $p-1$.

17. **[M26]** Не все значения, перечисленные в теореме D можно получить методом, изложенным в тексте. Например, число 11 не является первообразным элементом по модулю 5^e. Как это может быть если 11 (в соответствии с теоремой D)—первообразный элемент по модулю 10^e? Какие из чисел, перечисленных в теореме D,—первообразные элементы как по модулю 2^e, так и 5^e?

18. **[M25]** Докажите теорему D. (Ср. с предыдущим упражнением.)

19. **[40]** Составьте таблицу некоторых хороших множителей a для каждого из значений m, перечисленных в табл. 3.2.1.1-1, в предположении $c = 0$.

20. **[M24]** Какова длина периода линейной конгруэнтной последовательности, для которой (i) $X_0 = 0$; (ii) a—первообразный элемент по модулю $p_j^{e_j}$, $1 \le i \le t$; для всех степеней простых чисел в разложении $m = p_1^{e_1} \ldots p_t^{e_t}$ на простые множители; (iii) c и m взаимно просты?

3.2.1.3. Мощность. [6]

В предыдущем разделе мы показали, что максимальный период можно получить при $b = a - 1$, кратном всем простым делителям числа m (b также должно быть кратно 4, если m делится на 4). Если z—основание, которое используется в машине (так, $z = 2$—для двоичной машины и $z = 10$—для десятичной), а m—размер слова z^e машины, то множитель

$$a = z^k + 1, \qquad 2 \le k < e, \tag{1}$$

удовлетворяет этим условиям. Из теоремы 3.2.1.2.A также следует, что можно принять $c = 1$. Рекуррентное соотношение теперь имеет вид

$$X_{n+1} = ((z^k + 1)X_n + 1) \bmod z^e. \tag{2}$$

При вычислениях можно избежать умножения; достаточно простого сложения и сдвига.

Например, предположим, что $a = B^2 + 1$, где B—размер байта MIX. Вместо команд, приведенных в п.3.2.1.1, можно написать такую программу:

```
01 LDA   X
02 SLA   2
03 ADD   X
04 INCA  1
```
(3)

Время ее выполнения уменьшается с $16u$ до $7u$.

Ввиду сказанного множители вида (1) широко обсуждались в литературе и рекомендовались многими авторами. Однако почти пятилетние проверочные эксперименты показывают, что *следует избегать множителей, имеющих такой простой вид, как (1).* Причин здесь несколько. Прежде всего время счета не уменьшается на самом деле вдвое, как это происходит в примере (3). Если добавить к программе команды JMP, STJ, STA, JNOV, сравнительные времена счета будут равны $22u$ для мультипликативного метода и $13u$ для метода, использующего сложение и сдвиг. Кроме этого, необходимо учесть время работы основной программы, использующей случайные числа. Чистая экономия машинного времени в процентах почти ничтожна. А на многих современных машинах умножение выполняется *быстрее,* чем сдвиг и сложение!

Самый веский аргумент, препятствующий использованию множителя вида $z^k + 1$, заключается в том, что он приводит к недостаточно случайным числам. Одна из причин этого связана с концепцией "мощности", которую мы сейчас обсудим.

---

[6] В оригинале "potency",— *Прим. перев.*

Мощность линейной конгруэнтной последовательности с максимальным периодом определяется как наименьшее целое число s, такое, что

$$b^s \equiv 0 \pmod{m}. \tag{4}$$

(Такое целое s всегда существует, если множитель удовлетворяет условиям теоремы 3.2.1.2A, в частности, если b кратно любому простому делителю m.)

Мы можем анализировать случайность последовательности, принимая $X_0 = 0$, так как нуль обязательно встречается на протяжении ее периода. В таком случае имеем $X_n = ((a^n - 1)c/b) \bmod m$ и, разложив $a^n - 1 = (b+1)^n - 1$ по формуле бинома, находим

$$X_n = c\left(n + \binom{n}{2}b + \cdots + \binom{n}{s}b^{s-1}\right) \bmod m. \tag{5}$$

Все члены с b^s, b^{s+1} и т. д. можно опустить, так как они кратны m.

Исходя из уравнения (5), рассмотрим некоторые частные случаи. Если $a = 1$, мощность равна 1 и, как мы уже видели, $X_n \equiv cn \pmod{m}$, так что последовательность явно не случайная. Если мощность равна 2, имеем $X_n \equiv cn + cb\binom{n}{2}$, и снова последовательность нельзя считать случайной. Действительно,

$$X_{n+1} - X_n \equiv c + cbn,$$

так что разность между соседними случайными числами выражается очень простой зависимостью от n. Если

$$d = cb \bmod m,$$

точка (X_n, X_{n+1}, X_{n+2}) всегда лежит на одной из четырех плоскостей в трехмерном пространстве:

$$\begin{aligned} x - 2y + z &= d + m, \\ x - 2y + z &= d, \\ x - 2y + z &= d - m, \\ x - 2y + z &= d - 2m. \end{aligned}$$

Если мощность равна 3, последовательность выглядит несколько более случайной. Но X_n, X_{n+1}, и X_{n+2} все еще связаны сильной зависимостью. Разности $X_{n+1} - X_n$ образуют последовательность с мощностью 2, и тесты показывают, что последовательности с мощностью 3 все еще недостаточно хороши. Сообщалось, что приемлемые результаты могут быть получены для значения мощности, равного 4 и выше, но это оспаривалось многими авторами. Видимо, для достаточно случайных значений требуется мощность, равная по меньшей мере 5.

Предположим, например, что $m = 2$ и $a = 2^k + 1$. Тогда $b = 2^k$, так что при $k \geq 18$ значение $b^2 = 2^{2k}$ кратно m: мощность равна 2. Если $k = 17, 16, \ldots, 12$, мощность равна 3, а при $k = 11, 10, 9$ достигается значение 4. Поэтому с точки зрения мощности единственно приемлемы такие множители, для которых $k \leq 8$. Это значит, что $a \leq 257$, а мы увидим позже, что *небольших* множителей также следует избегать. Итак, все множители вида $2^k + 1$ при $m = 2^{35}$ оказываются неприемлемыми.

При больших размерах слова множители вида $2^k + 1$ принять можно. Был испытан и описан в литературе датчик с $m = 2^{47}$, $a = 2^9 + 1$ и мощностью, равной 6 (*CACM,* **4** (1961), 350–352). Несмотря на это, надо быть очень осторожным с множителями типа (1), потому что почти все известные ненадежные датчики были именно такого типа. В действительности даже приведенный пример не удовлетворяет статистическим тестам п.3.3.4.

При m, равном $w \pm 1$, где w—размер слова, m, вообще говоря, не разлагается на произведения высоких степеней простых чисел, так что большая мощность невозможна (см. пример 6). Поэтому в этом случае *не* стоит пользоваться методом максимального периода, а следует применять метод чистого умножения с $c = 0$.

Все еще остается много свободы в выборе множителя. Вообще говоря, мы хотим сохранить мощность высокой, множитель достаточно большим и, кроме того, избежать слишком простого по виду набора цифр в множителе. Предположим, что $m = 2^{35}$, а операция умножения ускоряется при уменьшении количества "единичных" битов в множителе. Можно рекомендовать (экспериментально) такой множитель, как $2^{23} + 2^{14} + 2^2 + 1$. Член 2^{23} делает множитель довольно большим. Член 2^2 обеспечивает высокую мощность. Единица необходима для получения максимального периода, а 2^{14} добавляется, чтобы множитель не оказался слишком простым для выработки достаточно случайной последовательности (ср. упр. 8). Член, подобный 2^{34}, был бы здесь не столь хорош, как 2^{23},

так как в произведении $2^{34}X_n$ используется только самый младший бит числа X_n (который не слишком случаен). Если скорость умножения не является лимитирующей, более "случайный" множитель (например, $a = 3141592621$), вероятно, окажется значительно более удовлетворительным.

В действительности концепция мощности дает только один из критериев выбора множителя; к нему можно добавить немало других. Ниже, в п.3.3.4, обсуждается "спектральный тест" для множителей линейных конгруэнтных последовательностей. Это—важный критерий, включающий, как частные случаи, мощность и величину множителя. В п.3.3.4 мы, например, увидим, что выбор $2^{23} + 2^{13} + 2^2 + 1$ намного хуже, чем $2^{23} + 2^{14} + 2^2 + 1$. Любой множитель, который будет широко использоваться, следует проверить спектральным тестом.

Упражнения

1. [M10] Покажите, что независимо от того, каким окажется размер байта B машины MIX, программа (3) служит датчиком случайных чисел с максимальным периодом.
2. [10] Какова мощность датчика, реализованного MIX-программой (3)?
3. [11] Какова мощность линейной конгруэнтной последовательности при $m = 2^{35}$, $a = 3141592621$? Чему равна мощность, если множитель $a = 2^{23} + 2^{13} + 2^2 + 1$?
4. [20] Покажите, что, если $m = 2^e \geq 8$, максимальная мощность достигается при $a \bmod 8 = 5$.
5. [M20] Дано, что $m = p_1^{e_1} \ldots p_t^{e_t}$ и $a = 1 + kp_1^{f_1} \ldots p_t^{f_t}$, где a удовлетворяет условиям теоремы 3.2.1.2A, а k и m взаимно просты. Покажите, что мощность равна $\max(\lceil e_1/f_1 \rceil, \ldots, \lceil e_t/f_t \rceil)$.

>6. [20] Какие из значений $m = w \pm 1$ в табл. 3.2.1.1-1 могут дать мощность, равную по меньшей мере 4? (Используйте результат упр. 5.)

7. [M20] Если число a удовлетворяет условиям теоремы 3.2.1.2A, оно взаимно просто с m; следовательно, существует число a', такое, что $aa' \equiv 1 \pmod{m}$. Покажите, что a' можно просто выразить с помощью b.

>8. [M26] Датчик случайных чисел с $X_{n+1} = (2^{17} + 3)X_n \bmod 2^{35}$ и $X_0 = 1$ подвергли следующему тесту. Пусть $Y_n = \lfloor 10X_n/2^{35} \rfloor$, тогда Y_n должна быть случайной цифрой между 6 и 9, а триады $(Y_{3n}, Y_{3n+1}, Y_{3n+2})$ должны принимать любое из 1000 возможных значений от $(0, 0, 0)$ до $(9, 9, 9)$ с равной вероятностью. Но в 30 000 проверенных чисел некоторые триады встречались очень редко, а некоторые появлялись гораздо чаще других. Можете ли вы объяснить такой странный результат?

3.2.2. Другие методы

Конечно, линейные конгруэнтные последовательности—не единственный из предложенных для вычислительных машин источников случайных чисел. В этом пункте мы приведем обзор других наиболее важных методов. Некоторые из них достаточно важны тогда как другие представляют интерес лишь постольку, поскольку оказываются совсем не такими хорошими, как кажутся на первый взгляд.

Одно из общепринятых заблуждений, с которыми приходится сталкиваться, когда речь идет о получении случайных чисел заключается в том, что достаточно взять хороший датчик и слегка его изменить, чтобы выработать "еще более случайную" последовательность. Довольно часто это неверно. Например, мы знаем что по формуле

$$X_{n+1} = (aX_n + c) \bmod m \tag{1}$$

можно получить довольно хорошие случайные числа Не будет ли последовательность

$$X_{n+1} = ((aX_n) \bmod (m + 1) + c) \bmod m \tag{2}$$

еще более случайной? Ответ таков, что новая последовательность с большей вероятностью *менее* случайна. Заметим, что целостная теория для нее рушится, а в отсутствие какой-либо теории о поведения последовательности (2) мы попадаем в область датчиков. типа $X_{n+1} = f(X_n)$ со случайно выбранной функцией f. Вместе с тем упр. 3.1-11–3.1-15 показывают, что эти последовательности ведут себя совсем не так хорошо, как если бы функция (1) была четко определена.

Рассмотрим другой подход, пытаясь генерировать "более случайные" числа. Линейный конгруэнтный метод можно обобщить, превратив его, скажем, в квадратичный конгруэнтный метод:

$$X_{n+1} = (dX_n^2 + aX_n + c) \bmod m. \tag{3}$$

В упр. 8 обобщается теорема 3.2.1.2A с целью получить необходимые и достаточные условия для a, c и d, такие, чтобы последовательность, определенная соотношением (3), имела бы максимальный период m. Ограничения оказываются не более жесткими, чем в линейном методе.

Для случая, когда m представляется степенью двойки, интересный квадратичный метод предложил Р. Ковэю. Пусть

$$X_0 \bmod 4 = 2, \quad X_{n+1} = X_n(X_n + 1) \bmod 2^e, \qquad n \geq 0. \tag{4}$$

Эту последовательность можно вычислять почти с той же эффективностью, как и (1), не заботясь о переполнении. Она имеет интересную связь с первоначальным методом середины квадрата фон Неймана. Возьмем Y_n, равное $2^e X_n$, так что Y_n—число, представленное с двойной точностью путем приписывания справа e нулей двоичному представлению X_n. Тогда Y_{n+1} состоит в точности из $2e$ средних цифр числа $Y_n^2 + 2^e Y_n$! Таким образом, метод Ковэю почти идентичен методу середины квадрата с двойной точностью, с той разницей, что он гарантирует большой период. Можно привести и дальнейшие доказательства случайности получаемой последовательности (см. упр. 3.3.4-25).

Другие обобщения соотношения (1) также довольно очевидны. Например, мы могли бы попытаться увеличить период последовательности. Период линейной конгруэнтной последовательности чрезвычайно велик. Обычно, если m приближается к размеру машинного слова, мы имеем дело с периодами порядка 10^9 и больше, так что в типичных задачах используется только очень малая часть последовательности. С другой стороны, величина периода влияет на степень случайности, которая достигается в последовательности (см. замечания, приведенные после соотношения 3.3.4-13). Поэтому обычно мы стремимся получать большой период, для чего и существует ряд методов. В одном из них вводится зависимость X_{n+1} от X_n и X_{n-1} вместо простой зависимости только от X_n. Тогда длину периода можно увеличить до m^2, так как последовательность начнет повторяться не раньше, чем будет выполнено равенство $(X_{n+\lambda}, X_{n+\lambda+1}) = (X_n, X_{n+1})$.

Простейший случай зависимости X_{n+1} от более чем одного из предыдущих значений реализуется в последовательности Фибоначчи

$$X_{n+1} = (X_n + X_{n-1}) \bmod m. \tag{5}$$

Этот датчик рассматривался в начале пятидесятых годов. Он дает обычно длину периода, большую, чем m. Однако тесты с определенностью показали, что числа, получаемые из соотношения Фибоначчи (5), являются *недостаточно* случайными. Поэтому в настоящее время формула (5) интересна главным образом как прекрасный "плохой пример".

Можно также рассмотреть датчики вида

$$X_{n+1} = (X_n + X_{n-k}) \bmod m, \tag{6}$$

где k—достаточно большое число, предложенные Грином, Смитом и Клемом (Green, Smith, Klem, *JACM,* **6** (1959), 527–537). При соответствующем выборе $X_0, X_1, \ldots, X_k$ эта формула обещает стать источником хороших случайных чисел. На первый взгляд соотношение (6) выглядит не слишком удобным для использования в машине, тем не менее существует очень эффективная процедура для ее реализации.

Алгоритм A. ***(Аддитивный датчик чисел.)*** Сначала в ячейки Z, $Y[0]$, $Y[1]$, $\ldots$, $Y[k]$ заносятся соответственно значения X_k, X_k, X_{k-1}, $\ldots$, X_0, а j принимается равным k. Последовательное использование алгоритма приводит к получению последовательности X_{k+1}, X_{k+2}, $\ldots$.

A1 **[$j < 0$?]** Если $j < 0$, установить $j \leftarrow k$.

A2 **[Прибавить, заменить.]** Установить $Z \leftarrow Y[j] \leftarrow (Z + Y[j]) \bmod m$.

A3 **[Уменьшить j.]** Уменьшить j на 1, выдать Z. ∎

Этот алгоритм на языке MIX выглядит так (при условии, что индексный регистр 6 не используется в основной программе):

```
01 J6NN  *+2     A1. j < 0?
02 ENT6  K       Установить j ← k.
03 LDA   Z       A2. Прибавить, заменить.
04 ADD   Y, 6    Z + Y[j] (возможно переполнение)
05 STA   Y, 6    → Y[j].
06 STA   Z       → Z.
07 DEC6  1       A3. Уменьшить j.
```
(7)

Этот датчик работает обычно быстрее, чем датчики, реализующие предыдущие методы, так как здесь не требуется никакого умножения.

Сейчас о таком аддитивном датчике известно немного. Прежде чем его можно будет рекомендовать, следует развить теорию, позволяющую установить необходимые показатели случайности вырабатываемых чисел, и провести широкие испытания для отдельных значений k и $X_0, X_1, \ldots, X_k$. Длина периода обсуждается в упр. 11; вообще говоря, она не намного больше m. В статье Грина, Смита и Клема говорится, что при $k \le 15$ последовательность не удовлетворяет тесту "проверка интервалов", описанному в п. 3.3.2, хотя при $k = 16$ тест проходит нормально.

Существует похожий, но гораздо более эффективный способ улучшения случайности линейных конгруэнтных последовательностей, если m—*простое число.* Например, m можно выбрать как самое большое простое число, которое можно записать в машинном слове. Такое число можно вычислить за приемлемое время, применяя технику п. 4.5.4. Когда $m = p$—простое число, из теории конечных полей следует, что существуют такие множители $a_1, \ldots, a_k$, что последовательность, определенная формулой

$$X_n = (a_1 X_{n-1} + \cdots + a_k X_{n-k}) \bmod p, \tag{8}$$

имеет период длины $p^k - 1$. Здесь $X_0, \ldots, X_{k-1}$ могут быть выбраны произвольно, но не должны быть все равны нулю. (Частный случай $k = 1$ соответствует мультипликативной конгруэнтной последовательности по простому модулю, с которой мы уже знакомы.) Выбор постоянных $a_1, \ldots, a_k$ в (8) тогда и только тогда дает желаемый результат, когда полином

$$f(x) = x^k - a_1 x^{k-1} - \cdots - a_k \tag{9}$$

является "*примитивным многочленом по модулю* p", т. е. имеет корень, являющийся *первообразным элементом поля из* p^k элементов[7] (см. упр. 4.6.2-16).

Конечно, простой факт *существования* подходящих констант $a_1, \ldots, a_k$, обеспечивающих длину периода $p^k - 1$, недостаточен для практических целей. Мы должны уметь *находить* их, не перебирая все p^k вариантов, так как p имеет порядок размера машинного слова. К счастью, существует в точности $\varphi(p^k - 1)/k$ подходящих комбинаций $(a_1, \ldots, a_k)$, так что имеется довольно большой шанс обнаружить одну из них после нескольких случайных попыток. Но, кроме всего прочего, нам нужно уметь быстро определить, является ли (9) примитивным многочленом по модулю p. Совершенно немыслимо вырабатывать $p^k - 1$ элементов последовательности в ожидании повторения! Методы проверки того, является ли многочлен примитивным по модулю p, обсуждались Эланеном и Кнутом (Alanen, Knuth, *Sankhyā*, Ser. A, **26** (1964), 305–328). Можно использовать следующий критерий.

Пусть $r = (p^k - 1)/(p - 1)$.

i) Величина $(-1)^{k+1} a_k$ должна быть первообразным корнем по модулю p (ср. п. 3.2.1.2).
ii) Полином x^r должен быть сравним с $(-1)^{k+1} a_k$ по модулю $f(x)$ и p.
iii) Степень $x^{r/q} \bmod f(x)$, где подразумевается полиномиальная арифметика по модулю p, должна быть положительной для всякого простого делителя q числа r.

Эффективные методы вычисления $x^n \bmod f(x)$, использующие полиномиальную арифметику по модулю простого p, обсуждаются в п. 4.6.2.

Чтобы сделать такую проверку, нам нужно знать факторизацию $r = (p^k - 1)/(p - 1)$ с помощью простых чисел. Это ограничивает возможности вычислений. За приемлемое время можно факторизовать r при $k = 2$, 3 и, может быть, 4, но с большими значениями k, если p велико, трудно иметь дело. Даже при $k = 2$ число "значимых случайных цифр" удваивается по сравнению с $k = 1$. Поэтому большие значения k редко необходимы.

Для оценки последовательности чисел, получаемых с помощью (8), можно воспользоваться вариантом спектрального теста, описанным в п. 3.3.4 (см. упр. 3.3.4-26). Из изложенного в этом пункте следует, что *нецелесообразно* ограничиваться очевидными значениями $a_1 = +1$ или $a_1 = -1$, даже если это возможно. Лучше взять большие, существенно "случайные" числа $a_1, \ldots, a_k$, удовлетворяющие условиям, а затем проверить выбор с помощью спектрального теста. Для определения $a_1, \ldots, a_k$, требуется провести много вычислений, но есть все основания считать, что в результате мы получим весьма удовлетворительный источник случайных чисел.

Особый интерес представляет значение $p = 2$. Иногда бывает нужен датчик, порождающий случайную последовательность *битов*—нулей и единиц (в отличие от дробей, принимающих значения от нуля до единицы). Существует простой способ вырабатывать на двоичной машине с k-разрядными словами весьма случайную последовательность битов. Начинаем с произвольного двоичного слова $\mathtt{Y} =$

[7] Этот элемент—образующая мультипликативной группы поля Галуа, которая, как известно, циклична.— *Прим. ред.*

$(Y_1Y_2\ldots Y_k)_2$, отличного от нуля. Чтобы получить очередной случайный бит последовательности, проделаем следующие операции, записанные на языке MIX:

```
01 LDA   Y      (Предполагаем, что сигнал переполнения выключен.)
02 ADD   Y      Сдвиг влево на один разряд.
03 JNOV  *+2    Переход, если в старшем разряде вначале был нуль.          (10)
04 XOR   C      В противном случае корректируем число операцией "исключающее или".
05 STA   Y
```

Четвертая по порядку операция, "исключающее или", имеется почти на всех двоичных машинах (ср. упр. 2.5-28). Она изменяет значение каждого разряда, соответствующего тому, где C содержит единицу, на обратное. В ячейке C находится двоичная константа $(a_1\ldots a_k)_2$, определяющая примитивный многочлен по модулю 2: $x^k - a_1x^{k-1} - \cdots - a_k$. После выполнения программы (10) в младшем разряде Y содержится следующий бит последовательности (если это более удобно, можно, наоборот, использовать старший разряд Y). Рассмотрим в качестве примера рис. 1, иллюстрирующий

1011
0101
1010
0111
1110
1111
1101
1001
0001
0010
0100
1000
0011
0110
1100
1011

Picture: Рис. 1. Последовательные состояния машинного слова Y при использовании двоичного метода для $k = 4$ и $c = \texttt{CONTENTS(C)} = (0011)_2$.

получение последовательности при $k = 4$, $c = (0011)_2$ (это, конечно, нестандартно малое значение k). Правый столбец таблицы представляет последовательность битов, которая повторяется с периодом $2^k - 1 = 15$: 1101011110001001 ... . Последовательность достаточно случайная, если учесть, что она получена с помощью четырех разрядов памяти. Чтобы убедиться в этом, рассмотрим соседние четверки битов, появляющиеся на протяжении периода, а именно: 1101, 1010, 0101, 1011, 0111, 1111, 1110, 1100, 1000, 0001, 0010, 0100, 1001, 0011, 0110. Вообще говоря, так как длина периода равна $2^k - 1$, каждая возможная комбинация k битов встречается за период точно один раз, за исключением набора из всех нулей. Таким образом, соседние наборы из k битов существенно независимы. В § 3.5 мы увидим, что существует очень мощный критерий случайности для k, равного, скажем, 30 или больше. Теоретические результаты, иллюстрирующие случайность этой последовательности, приводятся в статье Р. Таусворта (R. C. Tausworthe, *Math. Comp.*, **19** (1965), 201–209).

Примитивные многочлены степени ≤ 100 по модулю 2 были протабулированы Э. Уотсоном (E. J. Watson, *Math. Comp.*, **16** (1962), 368–369). При $k = 35$ можно принять

$$c = (00000000000000000000000000000000101)_2,$$

а для $k = 30$ можно взять

$$c = (00000000000000000000000001010011)_2.$$

Все же, как следует из упр. 18 и 3.3.4-26, для определения примитивных многочленов по модулю 2 лучше находить существенно "случайные" константы c.

Предупреждение: Некоторые попадались в ловушку, пытаясь использовать метод выработки случайных последовательностей битов для получения случайных дробей, занимающих целое слово $(.Y_0Y_1\ldots Y_{k-1})_2$, $(.Y_kY_{k+1}\ldots Y_{2k-1})_2$, ... На самом деле это довольно плохой способ, хотя отдельные биты каждой дроби вполне случайны (см. упр. 18)!

Мы уже видели, что, когда X_n определяется подходящей функцией от X_{n-1}, ..., X_{n-k}, можно найти такие последовательности с $0 \le X_n < m$ и периодом $m^k - 1$, где m—простое число. Наибольший период, который можно получить для *произвольной* последовательности, определенной соотношением

$$X_n = f(X_{n-1}, \ldots, X_{n-k}), \qquad 0 \le X_n < m, \tag{11}$$

как можно видеть, равен m^k. М. Мартин (M. H. Martin *Bull. Amer. Math. Soc.*, **40** (1934), 859–864) первый показал, что существуют функции, позволяющие достичь этого максимума для любых m и k. Его метод легко обосновать, но, к сожалению, он неудобен для программирования (см. упр. 17). Из известных функций f, дающих максимальный период m^k, самой простой является описанная в упр. 21. Соответствующие программы, вообще говоря, не так эффективны для выработки случайных чисел, как при реализации других ранее описанных методов. Все же они позволяют продемонстрировать явную случайность последовательности (когда речь идет о периоде в целом).

Другой важный класс методов сводится к *комбинации* датчиков случайных чисел для получения "еще более случайных" последовательностей. Всегда найдутся скептики, полагающие, что линейные конгруэнтные методы, аддитивные методы и т. д. слишком просты для выработки достаточно случайных последовательностей. А так как невозможно *доказать*, что их скептицизм неоправдан (хотя мы и верим, что это так), довольно бесполезно оспаривать подобное мнение. Существуют вполне эффективные методы для того, чтобы получать из двух последовательностей настолько случайную третью, что только самым отъявленным скептикам она может не понравиться.

Предположим, что мы имеем две последовательности X_0, X_1, ..., и Y_0, Y_1, ..., случайных чисел, расположенных между нулем и $m - 1$, полученные двумя независимыми способами. Одно из предложений сводится к тому, чтобы складывать числа попарно по модулю m, получая последовательность $Z_n = (X_n + Y_n) \bmod m$. В этом случае желательно, чтобы длины периодов $\langle X_n \rangle$ и $\langle Y_n \rangle$ были взаимно простыми числами (см. упр. 13).

Метод, предложенный Маклареном и Марсальей значительно лучше и удивительно удобен для программирования.

Алгоритм М. *(Вполне случайная последовательность.)* При заданных методах выработки двух последовательностей $\langle X_n \rangle$ и $\langle Y_n \rangle$ этот метод позволяет генерировать члены "значительно более случайной" последовательности. Мы используем вспомогательную таблицу $V[0]$, $V[1]$, ..., $V[k-1]$, где k—некоторое число, выбираемое обычно для удобства равным примерно 100. Сначала V-таблица заполняется первыми k значениями X-последовательности.

М1 **[Выработать X, Y.]** Установить в X и Y значения очередных членов последовательностей $\langle X_n \rangle$ и $\langle Y_n \rangle$ соответственно.

М2 **[Вычислить j.]** Установить $j \leftarrow \lfloor kY/m \rfloor$, где m—модуль, использующийся в последовательности $\langle Y_n \rangle$. Таким образом, j принимает случайное значение, определяемое с помощью Y; $0 \le j < k$.

М3 **[Заменить.]** Выдать $V[j]$ и установить $V[j] \leftarrow X$. ∎

Этот метод можно усиленно рекомендовать. Он позволяет получать невероятно большие периоды, если периоды последовательностей $\langle X_n \rangle$ и $\langle Y_n \rangle$ взаимно простые. И даже если длина периода не очень существенна, соседние члены последовательности почти не коррелируют друг с другом. Причиной того, что этот метод намного превосходит метод середины квадрата или метод, основанный на соотношении (2), является достаточная случайность последовательностей X_n и Y_n, которые не могут вырождаться. Читателю рекомендуется разобрать упр. 3, чтобы увидеть, как метод работает в частном случае.

На машине MIX можно реализовать алгоритм М, принимая k на единицу большим максимального значения, размещающегося в одном байте (равным размеру байта). Шаги М2 и М3 легко программируются следующим образом:

```
01 LD6   Y(1:1)     j ← старший байт Y.
02 LDA   V, 6       rA ← следующий элемент новой последовательности.
03 LDX   X
04 STX   V,6        V[j] ← X.
```
(12)

Для примера предположим, что алгоритм М применяется к таким двум последовательностям с $k = 64$:

$$X_0 = 5772156649, \quad X_{n+1} = (3141592653 X_n + 2718281829) \bmod 2^{35};$$
$$Y_0 = 1781072418, \quad Y_{n+1} = (2718281829 Y_n + 3141592653) \bmod 2^{35}.$$

Мы утверждаем, что последовательность, полученная с помощью алгоритма М, будет удовлетворять фактически *любому* критерию случайности для генерируемых вычислительной машиной последовательностей. Более того, время выработки чуть больше чем вдвое превышает время получения одной последовательности $< X_n >$.

Ф. Гебхардт показал [F. Gebhardt, *Math. Comp.*, **21** (1967),708–709], что алгоритм M позволяет получать удовлетворительные результаты, даже если его применять к таким неслучайным последовательностям, как последовательность Фибоначчи с $X_n = F_2 \bmod m$ и $Y_n = F_{2n+1} \bmod m$. Другой способ комбинировать две последовательности основан на циклическом сдвиге и "исключающем или" в двоичной машине. Его предложил У. Уэстлэйк (W. J. Westlake, *JACM*, **14** (1967), 337–340).

Упражнения

>1. [12] Практически мы получаем случайные числа, пользуясь соотношением $X_{n+1} = (aX_n + c) \bmod m$, где X_n—*целые*. После чего мы обращаемся с ними, как с *дробями:* $U_n = X_n/m$. Рекуррентная формула для U_n в действительности такова:

$$U_{n+1} = (aU_n + c/m) \bmod 1.$$

Обдумайте *прямое* использование этой формулы для выработки случайных последовательностей с помощью операций с плавающей точкой, имеющихся в машине.

>2. [M20] Для хорошего источника случайных чисел соотношение $X_{n-1} < X_{n+1} < X_n$ должно выполняться примерно в течение одной шестой всего времени, так как все возможные порядковые комбинации X_{n-1}, X_n и X_{n+1} равновероятны. Покажите, что указанный выше порядок тем не менее *никогда* не встречается в последовательности Фибоначчи (5).

3. [21] Какая последовательность получится при использовании алгоритма M, если $X_0 = 0$, $X_{n+1} = (5X_n + 3) \bmod 8$, $Y_0 = 0$, $Y_{n+1} = (5Y_n + 1) \bmod 8$ и $k = 4$? (Заметьте, что мощность равна двум, так что исходные $< X_n >$ и $\langle Y_n \rangle$ не слишком случайны.)

4. [00] Почему в первой команде программы (12) используется именно старший байт, а не какой-нибудь другой?

>5. [20] Рассмотрите возможность использования условия $X_n = Y_n$ для ускорения работы алгоритма M.

6. [10] В тексте при исследовании двоичного метода (10) утверждается, что младший бит слова X случаен, если многократно применять этот метод. Почему не случайно все *слово* X?

7. [20] Покажите, что можно получить полную последовательность длины 2^e (т. е. каждый из 2^e возможных вариантов соседних e битов, который реализуется только один раз на протяжении периода), если изменить программу (10) следующим образом:

```
01 LDA   X
02 JANZ  *+2
03 LDA   C
04 ADD   X
05 JNOV  *+3
06 JAZ   *+2
07 XOR   C
08 STA   X
```

8. [M39] Докажите, что квадратичная конгруэнтная последовательность (3) имеет период длины m тогда и только тогда, когда выполняются следующие условия:

i) c и m—взаимно простые числа;
ii) d и $a - 1$ кратны p—всем нечетным простым делителям m;
iii) d—четное и $d \equiv a - 1 \pmod 4$, если m кратно 4, $d \equiv a - 1 \pmod 2$, если m кратно 2;
iv) или $d = 0$, или $a \equiv 1$ и $cd \equiv 6 \pmod 9$, если m кратно 9. [*Указание.* Последовательность, определенная соотношениями $X_0 = 0$, $X_{n+1} = dX_n^2 + aX_n + c$, имеет по модулю m период длины m, если только эта длина периода равна d по модулю d, где d—произвольный делитель m.]

>9. [M24] (Р. Ковэю.) Используйте результат упр. 8, чтобы доказать, что в модифицированном методе середины квадрата (4) длина периода равна 2^{e-2}.

10. [M29] Покажите, что если X_0 и X_1 не являются оба четными и $m = 2^e$, то период последовательности Фибоначчи (5) равен $3 \cdot 2^{e-1}$.

11. [M36] Задача этого упражнения состоит в том, чтобы проанализировать определенные свойства целочисленных последовательностей, удовлетворяющих рекуррентному соотношению

$$X_n = a_1 X_{n-1} + \cdots + a_k X_{n-k}, \qquad n \geq k.$$

Если мы можем вычислить длину периода этой последовательности по модулю $m = p^e$, где p—простое число, то длина периода относительно произвольного модуля m равна наименьшему общему кратному длин периодов, вычисленных относительно степеней простых сомножителей m.

a) Пусть $f(z)$, $a(z)$, $b(z)$)—полиномы с целочисленными коэффициентами; будем писать $a(z) \equiv b(z) \pmod{f(z) \text{ и } m}$, если $a(z) = b(z) + f(z)u(z) + mv(z)$ для некоторых полиномов $u(z)$, $v(z)$ с целочисленными коэффициентами. Докажите, что при $f(0) = 1$ и $p^e > 2$ справедливо следующее: "если $z^\lambda \equiv 1 \pmod{f(z) \text{ и } p^e}$, $z^\lambda \not\equiv 1 \pmod{f(z) \text{ и } p^{e+1}}$, тогда $z^{p\lambda} \equiv 1 \pmod{f(z) \text{ и } p^{e+1}}$, $z^{p\lambda} \not\equiv 1 \pmod{f(z) \text{ и } p^{e+2}}$".

b) Пусть

$$f(z) = 1 - a_1 z - \cdots - a_k z^k,$$
$$G(z) = 1/f(z) = A_0 + A_1 z + A_2 z^2 + \ldots .$$

Обозначим символом $\lambda(m)$ длину периода последовательности $\langle A_n \bmod m \rangle$. Докажите, что $\lambda(m)$—наименьшее положительное целое λ, такое, что $z^\lambda \equiv 1 \pmod{f(z) \text{ и } m}$.

c) Пусть p—простое, $p^e > 2$ и $\lambda(p^e) \neq \lambda(p^{e+1})$. Докажите, что $\lambda(p^{e+r}) = p^r \lambda(p^e)$ для всех $r \geq 0$. (Таким образом, чтобы найти длину периода последовательности $\langle A_n \bmod 2^e \rangle$, можно вычислять $\lambda(4)$, $\lambda(8)$, $\lambda(16)$, ... вручную до тех пор, пока мы не найдем наименьшее $r \geq 2$, такое, что $\lambda(2^{r+1}) \neq \lambda(4)$. Тогда длина периода определена по $\bmod 2^e$ для всех e.)

d) Покажите, что любая последовательность целых чисел, удовлетворяющая рекуррентному соотношению, приведенному в начале упражнения, имеет производящую функцию $g(z)/f(z)$, где $g(z)$—некоторый полином с целочисленными коэффициентами.

e) Пусть полиномы $f(z)$ и $g(z)$ из (d) взаимно простые по модулю p (ср. п. 4.6.1). Докажите, что последовательность $\langle X_n \bmod p^e \rangle$ имеет длину периода в точности такую же, как и специальная последовательность $\langle A_n \bmod p^e \rangle$ в (b). (Никаким выбором $X_0, \ldots, X_{k-1}$ нельзя получить более длинный период, так как общая последовательность представляется линейной комбинацией "сдвигов" специальной последовательности.) [*Указание.* Существуют полиномы, такие, что $a(z)f(z) + b(z)g(z) \equiv 1 \pmod{p^e}$. Это следует из упр. 4.6.2-22 (лемма Гензеля).]

>12. [M28] Найдите целые числа X_0, X_1, a, b и c, такие, что последовательность

$$X_{n+1} = (aX_n + bX_{n-1} + c) \bmod 2^e, \qquad n \geq 1,$$

имеет самый большой период из всех последовательностей этого типа. [*Указание.* $X_{n+2} = ((a+1)X_{n+1} + (b-a)X_n - bX_{n-1}) \bmod 2^e$ См. упр. 11 (c).]

13. [M20] Пусть $\langle X_n \rangle$ и $\langle Y_n \rangle$— последовательности целых чисел по модулю m с периодами длины λ_1 и λ_2; образуем новую последовательность $Z_n = (X_n + Y_n) \bmod m$. Покажите, что, если λ_1 и λ_2—взаимно простые числа, последовательность $\langle Z_n \rangle$ имеет длину периода $\lambda_1 \lambda_2$.

14. [M24] Пусть X_n, Y_n, Z_n, λ_1, λ_2, такие же, как и в предыдущем упражнении. Предположим, что $\lambda_1 = 2^{e_2} 3^{e_3} 5^{e_5} \ldots$—разложение λ_1 на простые множители, и аналогично $\lambda_2 = 2^{f_2} 3^{f_3} 5^{f_5} \ldots$. Пусть $\lambda_0 = 2^{g_2} 3^{g_3} 5^{g_5} \ldots$, где $g_p = (\max(e_p, f_p)$, если $e_p \neq f_p$, и 0, если $e_p = f_p$). Покажите, что период λ' последовательности Z_n кратен λ_0, но является делителем λ—наименьшего общего кратного λ_1, λ_2. В частности, $\lambda' = \lambda$, если ($e_p \neq f_p$ есе $e_p = f_p = 0$) для всякого простого p.

15. [M46] Что можно сказать по поводу длины периода последовательности, вырабатываемой алгоритмом M?

>16. [M28] Пусть двоичное представление константы c, фигурирующей в методе (10), имеет вид $(a_1 a_2 \ldots a_k)_2$. Покажите, что последовательность битов Y_0, Y_1, ... удовлетворяет соотношению

$$Y_n = (a_1 Y_{n-1} + a_2 Y_{n-2} + \cdots + a_k Y_{n-k}) \bmod 2.$$

[Это можно рассматривать как другой способ определения последовательности. хотя на первый взгляд связь между этим соотношением и эффективной программой (10) не очевидна!]

17. [M33] (М. Мартин, 1934.) Пусть $m, k \geq 1$—целые числа и $X_1 = X_2 = \ldots = X_k = 0$. Для $n > 0$ положим X_{n+k} равным наибольшему неотрицательному значению $y < m$, такому, что k-набор $(X_{n+1}, \ldots, X_{n+k-1}, y)$ больше не встречается в последовательности. Другими словами, $(X_{n+1}, \ldots, X_{n+k-1}, y)$ не должен совпадать с $(X_{r+1}, \ldots, X_{r+k})$ для $0 \leq r < n$. В таком случае каждый возможный набор из k чисел будет встречаться в последовательности самое большее один раз. В конце концов процесс остановится, когда мы достигнем такого значения n, что $(X_{n+1}, \ldots, X_{n+k-1}, y)$ уже встречался в последовательности для всех y, $0 \leq y < m$. Например, если $m = k = 3$, последовательность такова:

$$000222122021121020120011101 00$$

и на этом процесс останавливается. a) Докажите, что, когда последовательность заканчивается, мы имеем $X_{n+1} = \ldots = X_{n+k-1} = 0$. b) Докажите, что *каждый* k-набор $(a_1, a_2, \ldots, a_k)$ элементов a_j, $0 \leq a_j < m$, встречается в последовательности. Следовательно, она заканчивается, когда $n = m^k$.

[*Указание.* Докажите индукцией по s, что появляется $(a_1, \ldots, a_s, 0, \ldots, 0)$ с $a_s \neq 0$.] Заметим, что если теперь определить $f(X_n, \ldots, X_{n+k-1}) = X_{n+k}$ для $1 \leq n \leq m^k$, полагая $X_{m^k+k} = 0$, то получим функцию с максимально возможным периодом.

18. [M22] Пусть Y_n — последовательность битов, полученная методом (10) при $k = 35$ и $c = (00000000000000000000000000000000101)_2$. Пусть U_n—двоичная дробь $(.Y_{nk}Y_{nk+1}\ldots Y_{nk+k-1})$. Покажите, что эта последовательность $\langle U_n \rangle$ не удовлетворяет тесту 3.3.2D при $d = 8$.
19. [M41] Найдите для каждого простого p из первого столбца табл. 1 в п. 4.5.4 подходящие (в смысле, указанном в тексте) константы a_1, a_2, такие, что длина периода (8) при $k = 2$ равна $p^2 - 1$.
20. [M40] Вычислите константы c, удобные для использования их в методе (10), имеющие примерно одинаковое число нулей и единиц, для $1 \leq k \leq 64$.
21. [M35] (Д. Рис.) В тексте объясняется, как находить функции f, такие, что у последовательности (11) длина периода равна $m^k - 1$ при условии, что m—простое число, а $X_0, \ldots, X_{k-1}$ отличны от нуля. Покажите, что эти функции можно модифицировать, чтобы получить последовательности вида (11) с длиной периода m^k для *всех* m. [*Указание.* Воспользуйтесь леммой 3.2.1.2Q, искусственным приемом упр. 7 и последовательностями вида $\langle pX_{2n} + X_{2n+1} \rangle$.]

>22. [M24] В тексте обсуждение обобщенных линейных последовательностей (8) ограничивается случаем, когда m—простое число. Докажите, что достаточно большие периоды можно получить, когда m "свободно от квадратов", т. е. представляется в виде произведения различных простых чисел. (Проверка табл. 3.2.1.1-1 показывает, что $m = w \pm 1$ часто удовлетворяет этой гипотезе. Многие результаты, полученные в тексте, можно поэтому применять и в этом случае, несколько более удобном для вычислений.)

3.3. СТАТИСТИЧЕСКИЕ ТЕСТЫ

Наша основная задача состоит в получении последовательностей, которые похожи на случайные. Мы уже видели, как добиться такого большого периода последовательности, чтобы в практических задачах исключить возможность ее повторения. Хотя это и важно, но большой период еще вовсе не означает, что последовательность хороша для работы. Как же решать, достаточно ли случайна последовательность?

Если дать любому человеку карандаш и бумагу и попросить его написать 100 случайных десятичных цифр, очень мало шансов на то, что он достаточно хорошо сможет с этим справиться. Люди стремятся избегать комбинаций, кажущихся им неслучайными, таких, как пары одинаковых соседних цифр (хотя примерно каждая из 10 цифр должна совпадать с предыдущей). Поэтому, увидев таблицу действительно случайных чисел, любой человек скорее всего скажет, что они совсем не случайные, его глаз сразу же отметит некоторые видимые закономерности.

Как заметил д-р Матрица (цитируется по работе M. Gardner, *Scientific American,* январь, 1965), "математики рассматривают десятичное представление числа π как случайный ряд, тогда как для современного толкователя чисел—это кладезь замечательных закономерностей". Д-р Матрица указал, например, что первое повторяющееся двузначное число в разложении π—это 26, а второе его появление приходится точно посередине одной любопытной конфигурации:

Picture: (1) p. 52

Выписав около дюжины других свойств этих цифр, он обнаружил, что, будучи правильно интерпретировано, число π отражает всю историю человечества!

Все мы выделяем особенности телефонных номеров, номерных знаков машин и т. д., чтобы легче их запомнить. Главная мысль всего сказанного заключается в том, что мы не можем доверять себе в оценке, случайна или нет данная последовательность чисел. Необходимо использовать какие-то непредвзятые механические тесты.

Статистическая теория дает нам некоторые количественные критерии случайности. Возможным же тестам буквально нет конца. Мы обсудим только те из них, которые, будучи наиболее полезными и поучительными, одновременно легко реализуются на вычислительных машинах.

Если последовательность ведет себя удовлетворительно относительно тестов $T_1, T_2, \ldots, T_n$, мы не можем быть *уверены* в том, что она выдержит, и следующее испытание T_{n+1}. Однако каждый тест дает нам все больше и больше уверенности в случайности последовательности. Обычно последовательность проверяется с помощью полудюжины разных тестов. Если их результаты оказываются удовлетворительными, мы считаем ее случайной (она считается невиновной до тех пор, пока не доказана ее виновность).

Каждую последовательность, которая будет интенсивно использоваться, следует тщательно проверить. Поэтому в следующих разделах объясняется, как правильно проводить такую проверку. Различаются два сорта тестов: *эмпирические тесты,* когда машина манипулирует с группами чисел последовательности и производит оценку с помощью определенных статистических критериев,

и *теоретические тесты,* когда мы находим некоторые характеристики последовательности, пользуясь методами теории чисел, базирующимися на рекуррентном соотношении, с помощью которого вырабатывается последовательность. В книге Д. Хаффа [D. Huff, How to Lie With Statistics, (Norton, 1954)] читатель может найти ряд других рекомендаций.

3.3.1. Универсальные тесты для анализа случайных последовательностей

А. Критерий χ^2. Критерий χ^2 ("хи-квадрат"), вероятно, самый распространенный из всех статистических критериев. Он используется не только сам по себе, но и как составная часть многих других тестов. Прежде чем приступить к общему описанию критерия χ^2, рассмотрим сначала в качестве примера, как можно было бы применить этот критерий для анализа игры в кости. Пусть каждый раз бросаются независимо две "правильные" кости, причем бросание каждой из них приводит с равной вероятностью к выпадению одного из чисел 1, 2, 3, 4, 5 и 6. Вероятности выпадения любой суммы s при одном бросании представлены в таблице:

Сумма	$s=$	2	3	4	5	6	7	8	9	10	11	12
Вероятность	$p_s=$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

(1)

(Например, сумма s=4 может быть получена тремя способами: $1+3$, $2+2$, $3+1$; при 36 возможных исходах это составляет $3/36=1/12=p_4$.)

Если бросать кости n раз, можно ожидать, что сумма s появится в среднем np_s раз. Например, при 144 бросаниях значение 4 должно появиться около 12 раз. Следующая таблица показывает, какие результаты были в *действительности,* получены при 144 бросаниях.

Сумма	$s=2$	3	4	5	6	7	8	9	10	11	12
Фактическое число выпадений	$Y_s=2$	4	10	12	22	29	21	15	14	9	6
Среднее число выпадений	$np_s=4$	8	12	16	20	24	20	16	12	8	4

(2)

Отметим, что фактическое число выпадений отличается от среднего во всех случаях. В этом нет ничего удивительного. Дело в том, что всего имеется 36144 возможных последовательностей исходов для 144 бросаний, и все они равновероятны. Одна из таких последовательностей состоит, например, только из двоек ("змеиные глаза"), и каждый, у кого "змеиные глаза" выпадут подряд 144 раза, будет уверен, что кости поддельные. Между тем эта последовательность так же вероятна, как и любая другая. Каким же образом в таком случае мы можем проверить, правильно ли изготовлена данная пара костей? Ответ заключается в том, что сказать определенно "да" или "нет" мы не можем, но можем дать *вероятностный* ответ, т. е. указать, насколько вероятно или невероятно данное событие.

Естественный путь решения нашей задачи состоит в следующем. Вычислим (прибегнув к помощи ЭВМ) сумму квадратов разностей фактического числа выпадений Y_s и среднего числа выпадений np_s (см. (2)):

$$V=(Y_2-np_2)^2+(Y_3-np_3)^2+\cdots+(Y_{12}-np_{12})^2. \tag{3}$$

Для плохого комплекта костей должны получаться относительно высокие значения V. Возникает вопрос, насколько вероятны такие высокие значения? Если вероятность их появления очень мала, скажем равна $1/100$,—т. е. отклонение результата от среднего значения на такую большую величину возможно только в одном случае из 100,—то у нас есть определенные основания для подозрений. (Не следует забывать, однако, что даже *хорошие* кости будут давать такое высокое значение V один раз из 100, так что для большей уверенности следовало бы повторить эксперимент и посмотреть, получится ли повторно высокое значение V.)

В статистику V все квадраты разностей входят с равным весом, хотя $(Y_7-np_7)^2$, например, вероятно, будет намного больше, чем $(Y_2-np_2)^2$, так как $s=7$ встречается в шесть раз чаще, чем $s=2$. Оказывается, что в "правильную" статистику, или по крайней мере такую, для которой доказано, что она наиболее значима, член $(Y_7-np_7)^2$ входит с множителем, который в шесть раз меньше множителя при $(Y_2-np_2)^2$. Таким образом, следует заменить (3) на следующую формулу:

$$V=\frac{(Y_2-np_2)^2}{np_2}+\frac{(Y_3-np_3)^2}{np_3}+\cdots+\frac{(Y_{12}-np_{12})^2}{np_{12}}. \tag{4}$$

Определенную таким образом величину V называют статистикой χ^2, соответствующей значениям $Y_2,\ldots,Y_{12}$, полученным в эксперименте. Подставляя в эту формулу значения из (2), получаем

$$V=\frac{(2-4)^2}{4}+\frac{(4-8)^2}{8}+\cdots+\frac{(9-8)^2}{8}+\frac{(6-4)^2}{4}=7\frac{7}{48}. \tag{5}$$

Теперь, естественно, возникает вопрос, является ли значение $7\frac{7}{48}$ настолько большим, что его случайное появление можно считать маловероятным. Прежде чем отвечать на этот вопрос, сформулируем критерий χ^2 в более общем виде.

Предположим, что все возможные результаты испытаний разделены на k категорий. Проводится n *независимых испытаний;* это означает, что исход каждого испытания абсолютно не влияет на исход остальных. Пусть p_s—вероятность того, что результат испытания попадет в категорию s, и пусть Y_s—число испытаний, которые действительно *попали* в категорию s. Сформируем статистику

$$V = \sum_{1\le s\le k} \frac{(Y_s - np_s)^2}{np_s}. \tag{6}$$

В предыдущем примере имелось 11 возможных исходов при каждом бросании костей, так что $k = 11$. **[Формулы (4) и (6) различаются только нумерацией: в одном случае она производится от 2 до 12, а в другом—от 1 до k.]**

Используя тождество $(Y_s - np_s)^2 = Y_s^2 - 2np_sY_s + n^2p_s^2$ и равенства

$$\begin{aligned} Y_1 + Y_2 + \cdots + Y_k &= n, \\ p_1 + p_2 + \cdots + p_k &= 1, \end{aligned} \tag{7}$$

можно преобразовать формулу **(6)** к виду

$$V = \frac{1}{n} \sum_{1\le s\le k} \left(\frac{Y_s^2}{p_s}\right) - n, \tag{8}$$

причем в большинстве случаев такая запись облегчает вычисления.

Вернемся к вопросу о том, какие значения V можно считать разумными. Ответ на это дает табл. 1, в которой приведено "распределение χ^2 с ν степенями свободы" при разных значениях ν. Следует пользоваться строкой таблицы с $\nu = k - 1$; *число "степеней свободы" равно* $k - 1$, *т. е. на единицу меньше числа категорий.*

Таблица 1

Некоторые данные для распределения χ^2

(Более полные таблицы см. в Handbook of Mathematical Functions, ed. by M. Abramowitz and I. A. Stegun, U. S. Government Printing Office, 1964, Table 26.8)

	$p = 99\%$	$p = 95\%$	$p = 75\%$	$p = 50\%$	$p = 25\%$	$p = 5\%$	$p = 1\%$
$\nu = 1$	0.00016	0.00393	0.1015	0.4549	1.323	3.841	6.635
$\nu = 2$	0.00201	0.1026	0.5753	1.386	2.773	5.991	9.210
$\nu = 3$	0.1148	0.3518	1.213	2.366	4.108	7.815	11, 34
$\nu = 4$	0.2971	0.7107	1.923	3.357	5.385	9.488	13.28
$\nu = 5$	0.5543	1.1455	2.675	4.351	6.626	11.07	15.09
$\nu = 6$	0.8720	1.635	3.455	5.348	7.841	12.59	16.81
$\nu = 7$	1.239	2.167	4.255	6.346	9.037	14.07	18.48
$\nu = 8$	1.646	2.733	5.071	7.344	10.22	15.51	20.09
$\nu = 9$	2.088	3.325	5.899	8.343	11.39	16.92	21.67
$\nu = 10$	2.558	3.940	6.737	9.342	12.55	18.31	23.21
$\nu = 11$	3.053	4.575	7.584	10.34	13.70	19.68	24.73
$\nu = 12$	3.571	5.226	8.438	11.34	14.84	21.03	26.22
$\nu = 15$	5.229	7.261	11.04	14.34	18.25	25.00	30.58
$\nu = 20$	8.260	10.85	15.45	19.34	23.83	31.41	37.57
$\nu = 30$	14.95	18.49	24.48	29.34	34.80	43.77	50.89
$\nu = 50$	29.71	34.76	42.94	49.33	56.33	67.50	76.15
$\nu > 30$	*приблизительно* $\nu + 2\sqrt{\nu}x_p + \frac{4}{3}x_p^2 - \frac{2}{3}$						
$x_p =$	-2.33	-1.64	$-.675$	0.00	0.675	1.64	2.33

(На интуитивном уровне это можно пояснить следующим образом: значения $Y_1, Y_2, \ldots, Y_k$ не совсем независимы, так как Y_1, согласно (7), можно вычислить, зная $Y_2, \ldots, Y_k$. Следовательно, имеется $k - 1$ степеней свободы. Более строгая аргументация будет приведена ниже.)

Если в таблице в строке ν и колонке p находится число x, то это означает, что значение V, определяемое по формуле (8), будет больше x с вероятностью p. Например, для $p = 5\%$ и $\nu = 10$ таблица дает значение $x = 18.31$; это означает, что $V > 18.31$ только в 5% всех случаев.

Предположим, что описанный процесс бросания костей моделируется на ЭВМ с помощью последовательности чисел, которые предполагаются случайными, и что получены следующие результаты:

	$s=2$	3	4	5	6	7	8	9	10	11	12
Эксперимент 1	$Y_s=4$	10	10	13	20	18	18	11	13	14	13
Эксперимент 2	$Y_s=3$	7	11	15	19	24	21	17	13	9	5

(9)

Вычисляя статистику критерия χ^2, получаем в первом случае $V_1 = 29\frac{59}{120}$, а во втором случае $V_2 = 1\frac{17}{120}$. Табличные значения, соответствующие 10 степеням свободы, показывают, что V_1 *явно слишком, велико;* V бывает больше, чем 23.2, только в одном проценте случаев! (Более полные таблицы показывают, что вероятность появления столь большого значения V равна 0.1%.) Таким образом, в эксперименте 1 зарегистрировано значительное отклонение от нормы.

С другой стороны, V_2 очень мало, потому что Y_s в эксперименте 2 оказались очень близки к средним значениям np_s [ср. с (2)]. Из таблицы распределения χ^2 следует, что в 99% случаев V должно быть больше, чем 2.56. Значение V_2 *явно слишком мало;* полученные в эксперименте значения V_3 настолько близки к среднем значениям, что невозможно считать этот эксперимент случайным испытанием. (В самом деле, из более полных таблиц следует, что при 10 степенях свободы такие низкие значения V встречаются только в 0.03% случаев). Наконец, с помощью таблицы распределения χ^2 можно проверить полученное нами в (5) значение $V = 7\frac{7}{48}$. Оно попадает в интервал между 75 и 50%, так что мы не можем считать его слишком высоким или слишком низким; данные, представленные в (2), удовлетворяют критерию χ^2.

Большим преимуществом рассматриваемого метода является то, что одни и те же табличные значения используются при любых n и любых вероятностях p_s. Единственной переменной является $\nu = k - 1$. На самом деле приведенные в таблице значение не являются абсолютно точными во всех случаях: *это приближенные значения, справедливые лишь при достаточно больших значениях* n. Как велико должно быть n? Достаточно большими можно считать такие значения n, при которых любое из np_s не меньше 5; однако лучше брать n значительно большими, чтобы повысить надежность критерия. Заметим, что в рассмотренных примерах мы брали $n = 144$, и np_2 равнялось всего 4, что противоречит только что сформулированному правилу. Единственная причина этого нарушения кроется в том, что автору надоело бросать кости; в результате числа из таблицы оказались не очень подходящими для нашего случая. Было бы гораздо лучше провести эти эксперименты на машине при $n = 1000$ или $10\,000$, или даже $100\,000$.

На самом деле вопрос о выборе n не так прост. Если бы кости были действительно неправильные, это проявлялось бы при сколь угодно больших n (см. упр. 12). Но при больших значениях n могут сглаживаться *локальные* отклонения, такие, как следующие друг за другом блоки чисел с сильным систематическим смещением в противоположные стороны. При действительном бросании костей этого можно не опасаться, так как все время используются одни и те же кости, но если речь идет о последовательности чисел, полученных на ЭВМ, то такой тип отклонения от случайного поведения вполне возможен. В связи

Picture: Рис. 2. Результаты 90 проверок с использованием критерия χ^2 (ср. с рис. 5).

с этим желательно проводить проверку с помощью критерия χ^2 при разных значениях n, но в любом случае эти значения должны быть довольно большими.

Итак, проверка с помощью критерия χ^2 заключается в следующем. Проводится n независимых испытаний, где n—достаточно большое число. (Следует избегать применения критерия χ^2 в случаях, если испытания не независимы; см., например, упр. 10, где рассмотрен случай, когда одна половина событий зависит от другой.) Подсчитывается число испытаний, результат которых относится к каждой из k категорий, и по формулам (6) или (8) вычисляется значение V. Затем V сравнивается с числами из табл. 1 при $\nu = k - 1$. Если V меньше значения, соответствующего $p = 99\%$, или больше значения, соответствующего $p = 1\%$, то результаты бракуются как недостаточно случайные. Если p лежит между 99 и 95% или между 5 и 1%, то результаты считаются "подозрительными"; при значениях p, полученных интерполяцией по таблице, заключенных между 95 и 90% или 10 и 5%, результаты "слегка подозрительны". Часто с помощью критерия χ^2 проверяют по крайней мере три раза разные части исследуемого ряда чисел, и, если не менее двух раз из трех результаты оказываются подозрительными, числа отбрасываются как недостаточно случайные.

Рассмотрим в качестве примера рис. 2, где схематически представлены результаты проверки с помощью критерия χ^2 шести последовательностей случайных чисел. Для каждой последовательности делалось пять разных проверок (основанных на критерии χ^2), каждая из которых повторялась

на трех разных участках последовательности. В датчике A использован метод Макларена-Марсальи (алгоритм 3.2.2M), в датчике E—метод Фибоначчи, остальные датчики соответствуют линейным конгруэнтным последовательностям со следующими параметрами:

Датчик B: $X_0 = 0$, $a = 3141592653$, $c = 2718281829$, $m = 2^{35}$.
Датчик C: $X_0 = 0$, $a = 2^7 + 1$, $c = 1$, $m = 2^{35}$.
Датчик D: $X_0 = 47594118$, $a = 23$, $c = 0$, $m = 10^8 + 1$.
Датчик F: $X_0 = 314159265$, $a = 2^{18} + 1$, $c = 1$, $m = 2^{35}$.

Результаты, приведенные на рис. 2, позволяют сделать следующие выводы. Датчики A, B, D прошли испытания удовлетворительно, датчик C находится на грани и должен быть, по-видимому, забракован, а датчики E и F определенно не прошли испытаний. Датчик F, безусловно, маломощен; датчики C и D обсуждались в литературе, но у них слишком мало значение a. В датчике D реализован метод вычетов в том виде, в каком он был впервые предложен Лемером в 1948 г., а в датчике C—линейный конгруэнтный метод с $c \neq 0$ также в его первоначальном виде (Ротенберг, 1960).

Несколько другой подход к суждению о результатах проверки по критерию χ^2, без использования таких понятий, как "подозрительный", "слегка подозрительный" и т. д., и менее позволяющий полагаться на мнение ad hoc, описывается ниже в этом разделе.

B. Критерий Колмогорова-Смирнова (КС-критерий). Как мы видели, критерий χ^2 применяется в тех случаях, когда результаты испытаний распадаются на конечное число k категорий. Однако нередко случайные величины могут принимать бесконечно много значений. В частности, бесконечно много значений принимают вещественные случайные числа в интервале между 0 и 1. Хотя множество значений случайных чисел, полученных в вычислительной машине, неизбежно ограничено, хотелось бы, чтобы это никак не сказывалось на результатах расчетов.

В теории вероятностей и статистике принято использовать одни и те же обозначения при описании дискретных и непрерывных распределений. Пусть требуется описать распределение значений случайной величины X. Это делается с помощью *функции распределения* $F(x)$, где

$$F(x) = \text{вероятность того, что } (X \le x).$$

На рис. 3 представлены три примера. Первый из них—функция распределения *случайного бита*, т. е. случайной величины X,

Picture: Рис. 3. Примеры функций распределения.

принимающей значения 0 или 1, каждое с вероятностью $1/2$. На рис. 3, b показана функция распределения *вещественной случайной величины, равномерно распределенной* между нулем и единицей, так что вероятность того, что $X \le x$, просто равна x, если $0 \le x \le 1$. Например, вероятность того, что $X \le \frac{2}{3}$, равна $\frac{2}{3}$. На рис. 3, c показано предельное распределение значений V в критерии χ^2 (при 10 степенях свободы); это же распределение, но в другой форме, было уже представлено в табл. 1. Заметим, что $F(x)$ всегда возрастает от 0 до 1 при увеличении x от $-\infty$ до $+\infty$.

Используя значения $X_1, X_2, \ldots, X_n$ случайной величины X, полученные в результате независимых испытаний, можно построить *эмпирическую функцию распределения* $F_n(x)$:

$$F_n(x) = \frac{\text{число таких } X_1, X_2, \ldots, X_n, \text{ которые} \le x}{n}. \tag{10}$$

На рис. 4 показаны три эмпирические функции распределения (вертикальные линии, строго говоря, не являются частью графика $F_n(x)$). Там же изображены и истинные функции распределения $F(x)$. При увеличении n функции $F_n(x)$ должны все более точно аппроксимировать $F(x)$.

Критерий Колмогорова—Смирнова (КС-критерий) можно использовать в тех случаях, когда функция $F(x)$ не имеет скачков. Он основан на *разности между* $F(x)$ *и* $F_n(x)$. Плохой датчик случайных чисел будет давать эмпирические функции распределения, плохо аппроксимирующие $F(x)$. На рис. 4, b приведен пример, когда значения X_i слишком велики, так что кривая эмпирической функции распределения проходит слишком низко. На рис. 4, c представлен еще худший случай; ясно, что такие большие расхождения между $F_n(x)$ и $F(x)$ крайне маловероятны; КС-критерий должен указать, насколько они маловероятны.

Для этого формируются следующие статистики:

$$\begin{aligned} K_n^+ &= \sqrt{n} \max_{-\infty<x<+\infty} (F_n(x) - F(x)); \\ K_n^- &= \sqrt{n} \max_{-\infty<x<+\infty} (F(x) - F_n(x)). \end{aligned} \tag{11}$$

Здесь K_n^+ показывает, каково максимальное отклонение для случая $F_n > F$, а K_n^-—каково максимальное отклонение для случая $F_n < F$. В примерах, показанных на рис. 4, значения этих статистик следующие:

$$\begin{array}{cccc} & a & b & c \\ K_{20}^+ & 0.492 & 0.134 & 0.313 \\ K_{20}^- & 0.536 & 1.027 & 2.101 \end{array} \tag{12}$$

[*Замечание.* Наличие множителя $\sqrt{n}$ в формуле (11) может показаться странным. Согласно упр. 6, при фиксированном x стандартное отклонение $F_n(x)$ от $F(x)$ пропорционально $1/\sqrt{n}$; следовательно, множитель $\sqrt{n}$ нормирует статистики K_n^+ и K_n^- таким образом, чтобы стандартное отклонение не зависело от n.]

Теперь можно, как и в случае критерия χ^2, найти значения K_n^+ и K_n^- в "процентильной" таблице, чтобы определить, имеют ли они высокую или низкую значимость. С этой целью для обеих величин K_n^+ и K_n^- можно воспользоваться табл. 2. Например, вероятность того, что K_{30}^- больше чем 0.8036, равна 25%. В отличие от таблицы для критерия χ^2, которой можно пользоваться только при больших n, в этой таблице приводятся точные значения

Picture: рис. 4. Примеры эмпирических распределений.

Таблица 2

Некоторые данные для распределения величин K_n^+ и K_n^-
(Для $n > 30$ приведены теоретически не обоснованные интерполяционные формулы, точные только при $n = \infty$)

	$p = 99\%$	$p = 95\%$	$p = 75\%$	$p = 50\%$	$p = 25\%$	$p = 5\%$	$p = 1\%$
$n = 1$	0.01000	0.05000	0.2500	0.5000	0.7500	0.9500	0.9900
$n = 2$	0.01400	0.06749	0.2929	0.5176	0.7071	1.0980	1.2728
$n = 3$	0.01699	0.07919	0.3112	0.5147	0.7539	1.1017	1.3589
$n = 4$	0.01943	0.08789	0.3202	0.5110	0.7642	1.1304	1.3777
$n = 5$	0.02152	0.09471	0.3249	0.5245	0.7674	1.1392	1.4024
$n = 6$	0.02336	0.1002	0.3272	0.5319	0.7703	1.1463	1.4144
$n = 7$	0.02501	0.1048	0.3280	0.5364	0.7755	1.1537	1.4246
$n = 8$	0.02650	0.1086	0.3280	0.5392	0.7797	1.1586	1.4327
$n = 9$	0.02786	0.1119	0.3274	0.5411	0.7825	1.1624	1.4388
$n = 10$	0.02912	0.1147	0.3297	0.5426	0.7845	1.1658	1.4440
$n = 11$	0.03028	0.1172	0.3330	0.5439	0.7863	1.1688	1.4484
$n = 12$	0.03137	0.1193	0.3357	0.5453	0.7880	1.1714	1.4521
$n = 15$	0.03424	0.1244	0.3412	0.5500	0.7926	1.1773	1.4606
$n = 20$	0.03807	0.1298	0.3461	0.5547	0.7975	1.1839	1.4698
$n = 30$	0.04354	0.1351	0.3509	0.5605	0.8036	1.1916	1.4801
$n > 30$	$\frac{0.07089}{\sqrt{n}} - \frac{0.15}{n}$	$\frac{0.1601}{\sqrt{n}} - \frac{0.14}{n}$	$\frac{0.3793}{\sqrt{n}} - \frac{0.15}{n}$	$\frac{0.5887}{\sqrt{n}} - \frac{0.15}{n}$	$\frac{0.8326}{\sqrt{n}} - \frac{0.16}{n}$	$\frac{1.2239}{\sqrt{n}} - \frac{0.17}{n}$	$\frac{1.5174}{\sqrt{n}} - \frac{0.20}{n}$

для всех n, так что КС-критерий можно применять при любом n.

Формулы (11) не годятся для машинных расчетов, так как требуется отыскать максимальное среди бесконечного множества чисел! Однако тот факт, что $F(x)$—неубывающая функция, а $F_n(x)$ имеет конечное число скачков, позволяет определить статистики K_n^+ и K_n^- с помощью следующего простого алгоритма:

Шаг 1. Определяются выборочные значения $X_1, X_2, \ldots, X_n$.

Шаг 2. Значения X_i располагаются в порядке возрастания так, чтобы $X_1 \le X_2 \le \ldots \le X_n$. (Эффективные алгоритмы, сортировки будут рассмотрены в гл. 5.)

Шаг 3. Нужные нам статистики вычисляются теперь по формулам

$$\begin{aligned} K_n^+ &= \sqrt{n} \max_{1 \le j \le n} \left(\frac{j}{n} - F(X_j) \right), \\ K_n^- &= \sqrt{n} \max_{1 \le j \le n} \left(F(X_j) - \frac{j-1}{n} \right). \end{aligned} \tag{13}$$

Сделать надлежащий выбор числа испытаний n в данном случае несколько легче, чем при работе с критерием χ^2, хотя некоторые трудности сохраняются. Если истинное распределение случайных величин X_j не отвечает функции $F(x)$, а описывается какой-то другой функцией $G(x)$, потребуется

сравнительно много испытаний, чтобы удостовериться, что $G(x) \neq F(x)$; n должно быть настолько большим, чтобы стало заметным различие между $G_n(x)$ и $F_n(x)$. С другой стороны, при больших n имеется тенденция к осреднению локальных отклонений от случайного поведения. Такие отклонения особенно нежелательны в большинстве приложений случайных чисел при работе на вычислительных машинах. С этой точки зрения было бы полезно *уменьшить* n. Тот факт что все n результатов испытаний надо запомнить, чтобы потом расположить в порядке возрастания, также склоняет нас к мысли уменьшить n. В качестве компромиссного решения можно взять n равным, скажем, 1000 и вычислить достаточно много значений K_{1000}^+ с использованием разных частей случайной последовательности:

$$K_{1000}^+(1), \quad K_{1000}^+(2), \quad \ldots, \quad K_{1000}^+(r). \tag{14}$$

К *этим* числам можно *опять* применить КС-критерий. Теперь уже $F(x)$—функция распределения для K_{1000}^+, причем она определяется эмпирической функцией распределения $F_r(x)$, построенной по случайным значениям (14). К счастью, функция распределения $F(x)$ очень проста: при больших значениях n, например $n = 1000$, распределение для K_n^+ хорошо аппроксимируется функцией

$$F_\infty(x) = 1 - e^{-2x^2}, \qquad x \geq 0. \tag{15}$$

Все сказанное относится и к K_n^-, так как K_n^+ и K_n^- ведут себя одинаково. *Такой подход, когда сначала критерий многократно применяется при относительно небольших* n*, а затем все результаты комбинируются, после чего КС-критерий применяется повторно, позволяет выявить как локальные, так и глобальные отклонения от заданного закона распределения.*

Полный эксперимент (хотя гораздо меньшего объема) был проведен автором в ходе работы над настоящей главой. Тест "наибольшее из 5", описанный в следующем разделе, применялся к 1000 равномерно распределенных случайных чисел. Получилось 200 чисел $X_1, X_2, \ldots, X_{200}$, относительно которых предполагалось, что они распределены в соответствии с функцией $F(x) = x^5$ ($0 \leq x \leq 1$). Все результаты были разделены на 20 групп по 10 чисел, и для каждой группы вычислялась статистика K_{10}^+. По полученным таким образом 20 значениям K_{10}^+ были построены эмпирические распределения, изображенные на рис. 4; гладкие кривые соответствуют предполагаемой функции распределения для статистики K_{10}^+. На рис. 4, a представлено эмпирическое распределение для K_{10}^+, соответствующее последовательности случайных чисел

$$\begin{aligned} Y_{n+1} &= (3141592653Y_n + 2718281829) \bmod 2^{35}, \\ U_n &= Y_n/2^{35}. \end{aligned}$$

Эту последовательность можно признать удовлетворительной. Рис. 4, b соответствует последовательности, полученной методом Фибоначчи; здесь наблюдаются отклонения *глобального* характера, т. е. можно показать, что значения X_n для теста "наибольшее из 5" не подчиняются распределению $F(x) = x^5$. На рис. 4, c приведены результаты проверки последовательности, пользующейся славой слабого алгоритма:

$$Y_{n+1} = ((2^{18} + 1)Y_n + 1) \bmod 2^{35}, \quad U_n = Y_n/2^{35}.$$

Результаты применения КС-критерия к этим данным были уже приведены в (12). Обращаясь к табл. 2 с $n = 20$, получаем, что значения K_{20}^+ и K_{20}^- в случае (b) слегка подозрительны (они попадают почти на 95- и 12%-ные уровни), но не настолько плохи, чтобы можно было с уверенностью отбросить эту последовательность. Значение K_{20}^- для случая (c), безусловно, никуда не годится, так что тест "наибольшее из 5" определенно доказывает непригодность этого датчика случайных чисел.

В описанном эксперименте КС-критерий должен выявлять глобальные отклонения от случайного поведения хуже, чем локальные, так как каждая группа состояла всего из 10 испытаний. Если бы было взято 20 групп по 1000 испытаний, отклонение в случае (b) было бы более значительным. Для иллюстрации этого КС-критерий был применен *сразу* ко всем данным, по которым построены графики на рис. 4. При этом получились следующие результаты:

	a	b	c
K_{200}^+	0.477	1.537	2.819
K_{200}^-	0.817	0.194	0.058

(16)

Теперь уже со всей определенностью выявились глобальные отклонения от заданного закона распределения, который дает метод Фибоначчи. Локальные отклонения в случае (c) сказываются вплоть до $n = 1\,000\,000$, так что при $n = 200$ говорить о глобальных отклонениях не приходится.

Критерий Колмогорова—Смирнова заключается, таким образом, в следующем. С помощью n *независимых испытаний* получаются значения $X_1, \ldots, X_n$ случайной величины с *непрерывной* функцией распределения $F(x)$. ($F(x)$ должна быть типа функций, изображенных на рис. 3, b и 3, c, т. е. без скачков как на рис. 3, a.) Затем с помощью процедуры, описанной перед формулой (13), вычисляются статистики K_n^+ и K_n^-. Эти статистики должны быть распределены в соответствии с табл. 2.

Теперь можно сравнить критерии Колмогорова—Смирнова и χ^2. Прежде всего следует заметить, что КС-критерием можно пользоваться *в сочетании с* критерием χ^2, чтобы получить лучшую процедуру, чем метод ad hoc, упомянутый при завершении описания критерия χ^2. (Это значит, что существует более совершенный способ, нежели проведение трех проверок, чтобы выяснить, как много результатов окажутся "подозрительными".) Предположим, что с помощью критерия χ^2 были независимо обработаны, скажем, 10 разных участков случайной последовательности, в результате чего были получены значения $V_1, V_2, \ldots, V_{10}$. Обычный подсчет количества подозрительно больших или малых значений V—не лучший способ анализа (хотя в экстремальных случаях он будет работать, и *очень* большие или *очень* малые значения могут служить указанием на то, что данная последовательность содержит слишком много локальных отклонений). Значительно лучше построить по этим 10 значениям эмпирическую функцию распределения и сравнить ее с теоретической функцией распределения, которую можно получить с помощью табл. 1. После определения статистик K_{10}^+ и K_{10}^- формируется окончательное суждение об исходе проверки последовательности с помощью критерия χ^2. При 10 или даже 100 значениях все это легко можно сделать вручную, используя графические методы; при большем числе значений V потребуется подпрограмма для машинного расчета распределения χ^2. Отметим, что *все 20 точек на рис. 4, с попадают в интервал между* 5- *и* 95%-*ными уровнями,* так что *каждая* из них в отдельности не может рассматриваться как подозрительная; однако суммарное эмпирическое распределение явно не совпадает с теоретическим.

Важное различие между критериями Колмогорова—Смирнова и χ^2 заключается в том, что первый из них применяется к распределениям, не имеющим скачков, в то время как второй—к кусочно постоянным распределениям (так как все результаты делятся на k категорий). Таким образом, области приложения этих критериев различны. Правда, критерий χ^2 можно применять и для непрерывных распределений, если разделить всю область определения $F(x)$ на k частей и пренебречь вариациями в пределах каждого интервала. Если, например, мы хотим выяснить, распределены ли значения $U_1, U_2, \ldots, U_n$ равномерно между нулем и единицей, т. е. соответствует ли их распределение функции $F(x) = x$ при $0 \le x \le 1$, будет естественно применить КС-критерий. Но можно также разделить интервал от 0 до 1 на $k = 100$ равных частей, подсчитать, сколько в каждую из них попадает значений U, после чего использовать критерий χ^2 с 99 степенями свободы. В настоящее время не существует достаточно определенных теоретических результатов, позволяющих сравнивать эффективности этих двух критериев. Автор обнаружил примеры, в которых КС-критерии выявлял отклонения от случайности более явственно, чем критерий χ^2, а также и примеры противоположного характера. Если, например, интервалы, на которые разбивается промежуток $(0, 1)$, пронумерованы от 0 до 99 и отклонения от средних значений положительны в интервалах 0–49 и отрицательны в интервалах 50–99, то эмпирическая функция распределения будет значительно дальше от $F(x)$, чем можно было бы предположить по значению χ^2. Но если отклонения положительны в интервалах 0, 2, ..., 98 и отрицательны в интервалах 1, 3, ..., 99, то эмпирическая функция распределения будет гораздо ближе к $F(x)$. Отсюда видно, что характер регистрируемых отклонений несколько различен. Для 200 чисел, по которым построены графики на рис. 4, критерий χ^2 с $k = 10$ дает значения V, равные 9.4, 17.7 и 39.3; в этом конкретном случае полученные значения вполне сравнимы с теми, которые дает КС-критерий (см. (16)). Для непрерывных распределений КС-критерий обладает определенными преимуществами, так как критерий χ^2 по определению имеет приближенный характер и требует относительно больших значений n.

Интересно также посмотреть, как изменятся результаты проверки шести разных датчиков случайных чисел, приведенные на рис. 2, если вместо критерия χ^2 использовать КС-критерий. Эти результаты были получены для $n = 200$ с помощью критерия "наибольшее из t" при $1 \le t \le 5$; промежуток $(0, 1)$ разделялся на 10 равных частей. По ним можно вычислить статистики K_{200}^+ и K_{200}^- и полученные результаты представить в том же виде, как на рис. 2 (отмечая, какие значения выходят за уровень 99% и т. д.). Это сделано на рис. 5. Отметим, что датчик D (метод Лемера), судя по рис. 5, весьма плох, в то время как результаты обработки *тех же самых данных* по критерию χ^2 не обнаружили никаких отклонений. Датчик E (метод Фибоначчи), наоборот, на рис. 5 выглядит несколько лучше. Хорошие датчики A и B удовлетворяют всем критериям. Причины расхождений между рис. 2 и 5 заключаются в первую очередь в том, что (a) число испытаний 200 на самом деле недостаточно ве-

лико; (b) разделение результатов на "негодные", "подозрительные" и "слегка подозрительные" само по себе довольно подозрительно.

(Было бы несправедливостью обвинять Лемера в том, что он использовал в 1948 г. "плохой" датчик случайных чисел, потому что в данном конкретном случае это было вполне законно. Машина ENIAC работала в параллельном режиме и программировалась посредством коммутационной панели. Лемер установил такой режим, при котором содержимое одного из сумматоров

Picture: Рис. 5. Результаты применения КС-критерия к тем же данным, что и на рис. 2.

постоянно умножалось на 23 по модулю 10^8+1; содержимое сумматора изменялось каждые несколько микросекунд. Множитель 23 слишком мал, чтобы такую последовательность можно было считать случайной: корреляция между следующими непосредственно друг за другом числами при таком множителе слишком велика. Но промежуток времени между обращениями к сумматору, содержащему случайное число, был сравнительно велик и, кроме того, все время менялся. Так что фактически множитель был равен 23^k, где k—большое, изменяющееся число!)

C. История, библиография и теория. Критерий χ^2 был предложен Карлом Пирсоном в 1900 г. (*Philosophical Magazine,* Series 5, **50**, 157–175). Эта работа Пирсона считается одной из основополагающих в современной статистике, так как до нее качество экспериментальных результатов определялось просто по тому, как они выглядят на графике. В своей статье Пирсон приводит несколько интересных примеров неправильного использования статистики. Он показал также, что в некоторых случаях рулетка (на которой он экспериментировал в Монте-Карло в течение двух недель в 1892 г.) давала результаты, настолько далекие от средних значений, что при правильном устройстве рулетки они могли бы осуществиться не чаще, чем один раз из 10^{29}. Общее обсуждение критерия χ^2 и обширная библиография содержатся в обзорной статье У. Кокрэна (W. G. Cochran, *Annals of Mathematical Statistics,* **23** (1952), 315–345).

Изложим в сокращенном виде теорию, лежащую в основе критерия χ^2. Точная вероятность того, что $Y_1 = y_1, \ldots, Y_k = y_k$ равна, как легко видеть,

$$\frac{n!}{y_1! \ldots y_k!} p_1^{y_1} \ldots p_k^{y_k}. \tag{17}$$

Если предположить, что Y_s имеет распределение Пуассона с плотностью

$$\frac{e^{-np_s}(np_s)^{y_s}}{y_s!}$$

и случайные величины Y_s независимы, то вероятность осуществления значений $(y_1, \ldots, y_k)$ равна

$$\prod_{1\le s\le k} \frac{e^{-np_s}(np_s)^{y_s}}{y_s!},$$

а сумма $Y_1 + \cdots + Y_k$ будет равна n с вероятностью

$$\sum_{\substack{y_1+\cdots+y_k=n \\ y_1,\ldots,y_k\ge 0}} \prod_{1\le s\le k} \frac{e^{-np_s}(np_s)^{y_s}}{y_s!} = \frac{e^{-n}n^n}{n!}.$$

Если потребовать соблюдения условия $Y_1 + \cdots + Y_k = n$, а в *остальном* считать случайные величины независимыми, то вероятность того, что $(Y_1, \ldots, Y_k) = (y_1, \ldots, y_k)$, будет равна

$$\left(\prod_{1\le s\le k} \frac{e^{np_s}(np_s)^{y_s}}{y_s!}\right) \Big/ \left(\frac{e^{-n}n^n}{n!}\right),$$

что совпадает с (17). *Можно, следовательно, считать, что случайные величины Y имеют распределение Пуассона и независимы, с единственным ограничением, что их сумма фиксирована.*

Обозначим

$$Z_s = \frac{Y_s - np_s}{\sqrt{np_s}}, \quad V = Z_1^2 + \cdots + Z_k^2. \tag{18}$$

Условие $Y_1 + \cdots + Y_k = n$ можно записать в виде

$$\sqrt{p_1}Z_1 + \cdots + \sqrt{p_k}Z_k = 0. \tag{19}$$

Рассмотрим $(k-1)$-мерное пространство S векторов $(Z_1, \ldots, Z_k)$, для которых выполняется условие (19). При больших значениях n случайные величины Z_s приблизительно нормальны (см. упр. 1.2.10-16), так что вероятность выбора точки, принадлежащей элементарному объему $dZ_2 \ldots dZ_k$ в S *приближенно* пропорциональна $\exp(-(Z_1^2 + \cdots + Z_k^2)/2)$. (Именно этот момент в выводе приводит к тому, что критерий χ^2 является лишь аппроксимацией, справедливой для больших n.) Теперь вероятность того, что $V \le v$, равна

$$\frac{\displaystyle\int_{\substack{(Z_1,\ldots,Z_k)\ \text{в}\ S \\ \text{и}\ Z_1^2+\cdots+Z_k^2\le v}} \exp(-(Z_1^2 + \cdots + Z_k^2)/2)\, dz_2 \ldots dz_k}{\displaystyle\int_{(Z_1,\ldots,Z_k)\ \text{в}\ S} \exp(-(Z_1^2 + \cdots + Z_k^2)/2)\, dz_2 \ldots dz_k}. \tag{20}$$

Так как плоскость (19) проходит через начало координат в k-мерном пространстве, интегрирование в знаменателе проводится по объему сферы в $(k-1)$-мерном пространстве с центром в начале координат. Преобразованием к обобщенным полярным координатам с радиусом χ и углами $\omega_1, \ldots, \omega_{k-2}$ формула (20) приводится к виду

$$\frac{\displaystyle\int_{\chi^2\le v} e^{-\chi^2/2}\chi^{k-2} f(\omega_1, \ldots, \omega_{k-2})\, d\chi\, d\omega_1 \ldots d\omega_{k-2}}{\displaystyle\int e^{-\chi^2/2}\chi^{k-2} f(\omega_1, \ldots, \omega_{k-2})\, d\chi\, d\omega_1 \ldots d\omega_{k-2}};$$

функция f определена в упр. 15. При интегрировании по углам $\omega_1, \ldots, \omega_{k-2}$ в числителе и знаменателе появляется множитель, на который можно сократить. В конце концов получается формула

$$\frac{\displaystyle\int_0^{\sqrt{v}} e^{-\chi^2/2}\chi^{k-2}\, d\chi}{\displaystyle\int_0^{\infty} e^{-\chi^2/2}\chi^{k-2}\, d\chi}, \tag{21}$$

приближенно представляющая вероятность того, что $V \le v$.

В этом выводе радиус обозначен через χ, как в оригинальной работе Пирсона; отсюда критерий χ^2 и получил свое название. Подставляя $t = \chi^2/2$, можно выразить интегралы через неполную гамма-функцию, которая уже использовалась в п. 1.2.11.3:

$$\lim_{n\to\infty} P\{V \le v\} = \gamma\left(\frac{k-1}{2}, \frac{v}{2}\right) \Big/ \Gamma\left(\frac{k-1}{2}\right). \tag{22}$$

Таково определение распределения χ^2 с $(k-1)$ степенями свободы.

Перейдем теперь к КС-критерию. В 1933 г. А. Н. Колмогоров предложил критерий, основанный на статистике

$$K_n = \sqrt{n} \max_{-\infty<x<+\infty} |F_n(x) - F(x)| = \max(K_n^+, K_n^-). \tag{23}$$

Среди ряда модификаций этого критерия, описанных в 1939 г. Н. В. Смирновым, была и та, которая используется в этой книге и опирается на раздельное вычисление K_n^+ и K_n^-. Существует обширное семейство разновидностей этого критерия, но те из них, которые используют K_n^+ и K_n^-, особенно удобны для машинных расчетов. Исчерпывающий обзор литературы по КС-критерию и его обобщениям, включая обширный список литературы, содержится в работе Дарлинга (*Annals of Mathematical Statistics,* **28** (1957), 823–838). Работа проф. Дарлинга предназначена главным образом для специалистов в данной области.

Прежде чем начать изучение распределения K_n^+ и K_n^-, сформулируем следующее фундаментальное утверждение. *Если X—случайная величина с непрерывной функцией распределения $F(x)$, то $F(X)$—случайное число, равномерно распределенное между* 0 *и* 1. Для доказательства этого достаточно проверить, что при $0 \le y \le 1$ неравенство $F(X) \le y$ справедливо с вероятностью y. Из непрерывности $F(x)$ следует, что $F(x_0) = y$ для какого-то x_0. Поэтому вероятность того, что $F(X) \le y$,

есть вероятность выполнения неравенства $X \le x_0$. По определению эта вероятность есть $F(x_0)$, т. е. равна y.

Пусть $Y_j = nF(X_j)$ для $1 \le j \le n$. Тогда величины Y_j распределены равномерно между 0 и n; они независимы, за исключением того, что $Y_1 \le Y_2 \le \ldots \le Y_n$ (это эквивалентно требованию $X_1 \le X_2 \le \ldots \le X_n$, которое принимается при расчетах по формуле (13)). Формулу (13) можно переписать теперь следующим образом:

$$K_n^+ = \frac{1}{\sqrt{n}} \max(1 - Y_1, 2 - Y_2, \ldots, n - Y_n).$$

Если $0 \le t \le n$, то вероятность неравенства $K_n^+ \le t/\sqrt{n}$ есть вероятность того, что $Y_j \ge j - t$ для $1 \le j \le n$. Эту вероятность нетрудно записать в виде n-кратного интеграла

$$\frac{\displaystyle\int_{\alpha_n}^{n} dY_n \int_{\alpha_{n-1}}^{Y_n} dY_{n-1} \ldots \int_{\alpha_1}^{Y_2} dY_1}{\displaystyle\int_{0}^{n} dY_n \int_{0}^{Y_n} dY_{n-1} \ldots \int_{0}^{Y_2} dY_1}, \qquad \text{где } \alpha_j = \max(j - t, 0). \tag{24}$$

Знаменатель этого выражения равен $n^n/n!$. Действительно, куб, построенный на векторах $(Y_1, Y_2, \ldots, Y_n)$ с $0 \le Y_j < n$, имеет объем n^n; его надо разделить на $n!$ равных частей, соответствующих всем возможным упорядочениям Y_j. Интеграл в числителе несколько сложнее, но его можно взять способом, описанным в упр. 17; в результате получается общая формула

$$P\left\{K_n^+ \le \frac{t}{\sqrt{n}}\right\} = \frac{t}{n^n} \sum_{0 \le k \le t} \binom{n}{k} (k - t)^k (t + n - k)^{n-k-1}. \tag{25}$$

Распределение K_n^- точно такое же. Формула (25) была впервые получена Бирнбаумом и Тинги; ее можно использовать в случаях, которые не охвачены в табл. 2.

В оригинальной работе Смирнова показано, что

$$\lim_{n\to\infty} P\{K_n^+ \le s\} = 1 - e^{-2s^2} \qquad \text{при } s \ge 0. \tag{26}$$

В сочетании с формулой (25) это дает

$$\lim_{n\to\infty} \frac{s}{\sqrt{n}} \sum_{\sqrt{n}s < k \le n} \binom{n}{k} \left(\frac{k}{n} - \frac{s}{\sqrt{n}}\right)^k \left(\frac{s}{\sqrt{n}} + 1 - \frac{k}{n}\right)^{n-k-1} = e^{-2s^2}, \qquad s \ge 0. \tag{27}$$

Но доказать это предельное соотношение очень трудно. Наиболее элементарное доказательство принадлежит Уиттлу (*Annals of Mathematical Statistics,* **32** (1961), 499–505). Оно основано на следующем приеме. В результате некоторого увеличения и уменьшения α в (24) вводятся специальным образом модифицированные статистики, для которых вероятности вычисляются точно, и затем показывается, что как верхняя, так и нижняя границы стремятся к e^{-2s^2}.

Упражнения

1. [00] Какую строку из таблицы распределения χ^2 надо использовать для выяснения, не слишком ли велико значение $V = 7\frac{7}{48}$ в (5)?
2. [20] Определить вероятность p_s появления суммы s для $2 \le s \le 12$ при бросании двух костей, устроенных следующим образом: в одной из них 1, а в другой 6 выпадают ровно вдвое чаще, чем любое другое значение.

>3. [23] Кости, описанные в предыдущем упражнении, бросались 144 раза. Были получены следующие результаты:

s	2	3	4	5	6	7	8	9	10	11	12
Y_s	2	6	10	16	18	32	20	13	16	9	2

Примените к *этим* значениям критерий χ^2, пользуясь вероятностями (1), т. е. предполагая, что кости правильные. Позволит ли в таком случае критерий χ^2 определить, что кости неправильные? Если нет, то объясните почему.

>4. **[23]** На самом деле автор получил результаты (9) в эксперименте 1, имитируя кости, одна из которых нормальная, а на другой выпадает только 1 или 6 с равными вероятностями. Вычислите для этого случая вероятности, которые заменят (1), и с помощью критерия χ^2 определите, соответствуют ли результаты эксперимента такому устройству костей.

5. **[22]** Пусть $F(x)$—равномерное распределение, изображенное на рис. 3,b. Вычислите K_{20}^+ и K_{20}^- для следующих 20 экспериментальных значений:

$$\begin{array}{cccccccccc} 0.414, & 0.732, & 0.236, & 0.162, & 0.259, & 0.442, & 0.189, & 0.693, & 0.098, & 0,302, \\ 0.442, & 0.434, & 0.141, & 0.017, & 0.318, & 0.869, & 0.772, & 0.678, & 0.354, & 0.718. \end{array}$$

Определите с помощью каждого из этих двух критериев, соответствуют ли экспериментальные данные равномерному распределению.

6. **[M20]** Рассмотрим функцию $F_n(x)$, определенную формулой (10), при фиксированном x. Какова вероятность, что $F_n(x) = s/n$ для заданного целого s? Каково среднее значение $F_n(x)$? Каково среднеквадратичное отклонение?

7. **[M15]** Покажите, что K_n^+ и K_n^- не могут быть отрицательными. Каково наибольшее возможное значение K_n^+?

8. **[00]** В тексте был описан эксперимент, в котором в результате обработки последовательности случайных чисел было получено 20 значений статистики K_{10}^+. Эти данные, изображенные на рис. 4, были исследованы с помощью КС-критерия. Почему при этом использовались табличные значения для $n = 20$, а не для $n = 10$?

>9. **[20]** В описанном в тексте эксперименте на график наносились 20 значений K_{10}^+, вычисленных с помощью теста "наибольшее из 5", который применялся к разным частям случайной последовательности. Можно было бы вычислить также соответствующие 20 значений K_{10}^-; так как распределения статистик K_{10}^+ и K_{10}^- совпадают, можно было бы объединить все 40 значений (20 значений K_{10}^+ и 20 значений K_{10}^- и применить к ним КС-критерий, вычислив K_{40}^+ и K_{40}^-. Обсудите достоинства этой идеи.

>10. **[20]** Предположим, что результаты n испытаний обработаны с помощью критерия χ^2 и получено значение V. Затем те же n результатов обрабатываются еще раз (получается, естественно, то же самое), после чего данные обоих тестов объединяются и рассматриваются как единый тест с числом испытаний $2n$ (при этом нарушается требование независимости испытаний). Какая связь имеется между вторым и первым значением V?

11. **[20]** Решите упр. 10, заменив критерий χ^2 на критерий КС.

12. **[M26]** Пусть при применении критерия χ^2 к результатам n испытаний предполагается, что p_s— вероятность попадания в категорию s. Допустим, что в действительности вероятность попадания в категорию s равна $q_s \neq p_s$ (см. упр. 3). Конечно, желательно, чтобы критерий χ^2 позволил обнаружить, что первоначальное предположение было неверным. Покажите, что при достаточно больших n это действительно будет обнаружено. Докажите также аналогичный результат для КС-критерия.

13. **[M24]** Покажите, что формула (13) эквивалентна формуле (11).

>14. **[BM26]** Пусть величины Z_s заданы выражением (18). Покажите непосредственно с помощью формулы Стерлинга, что справедливо следующее равенство для полиномиального распределения:

$$n' p_1^{Y_1} \dots p_k^{Y_k} / Y_1! \dots Y_k! = e^{-V/2} / \sqrt{(2n\pi)^{k-1} p_1 \dots p_k} + O(n^{-k/2}),$$

если Z_1, Z_2, $\dots$, Z_k ограничены при $n \to \infty$. (Такой подход позволяет обосновать критерий χ^2, больше опираясь на первоосновы и проводя меньше выкладок, чем при подходе, использованном в тексте.)

15. **[BM24]** Обычно полярные координаты в двумерном пространстве задаются выражениями $x = r\cos\theta$ и $y = r\sin\theta$. При интегрировании используется равенство $dx\,dy = r\,dr\,d\theta$. В общем случае для n-мерного пространства можно положить

$$\begin{aligned} x_k &= r\sin\theta_1 \dots \sin\theta_{k-1}\cos\theta_k, \qquad 1 \le k < n, \\ x_n &= r\sin\theta_1 \dots \sin\theta_{n-1}. \end{aligned}$$

Покажите, что в этом случае

$$dx_1\,dx_2\,\dots\,dx_n = \left| r^{n-1} \sin^{n-2}\theta_1 \dots \sin\theta_{n-2}\,dr\,d\theta_1\,\dots\,d\theta_{n-1} \right|.$$

16. **[BM35]** Обобщите теорему 1.2.11.3A так, чтобы определить значение

$$\gamma(x+1, x+z\sqrt{2x}+y)/\Gamma(x+1)$$

при больших x и фиксированных y, z. Опустите члены $O(1/x)$. С помощью полученного результата найдите приближенное решение t уравнения

$$\gamma\left(\frac{\nu}{2}, \frac{t}{2}\right) \Big/ \Gamma\left(\frac{\nu}{2}\right) = p$$

при больших ν и фиксированном p; таким образом будет найдено объяснение асимптотических формул, приведенных в табл. 1. (*Указание:* см. упр. 1.2.11.3-8.)

17. [BM26] Пусть t—некоторое действительное число, и пусть при $0 \le k \le n$

$$P_{nk}(x) = \int_{n-t}^{x} dx_n \int_{n-1-t}^{x_n} dx_{n-1} \ldots \int_{k+1-t}^{x_{k+2}} dx_{k+1} \int_{0}^{x_k+1} dx_k \ldots \int_{0}^{x_2} dx_1,$$

а $P_{00}(x) = 1$. Докажите следующие соотношения:

a) $P_{nk}(x) = \displaystyle\int_{n}^{x+t} dx_n \int_{n-1}^{x_n} dx_{n-1} \ldots \int_{k+1}^{x_{k+2}} dx_{k+1} \int_{t}^{x_k+1} dx_k \ldots \int_{x}^{x_2} dx_1$;

b) $P_{n0}(x) = (x+t)^n/n! - (x+t)^{n-1}/(n-1)!$;

c) $P_{nk}(x) - P_{n(k-1)}(x) = \dfrac{(k-t)^k}{k!} P_{(n-k)0}(x-k)$ при $1 \le k \le n$;

d) получите общую формулу для $P_{nk}(x)$ и используйте ее для вывода формулы (24).

18. [M20] Дайте "простое" объяснение, почему распределения величин K_n^+ и K_n^- одинаковы.
19. [BM48] Разработайте критерии, подобные критерию Колмогорова-Смирнова, для многомерных распределений $F(x_1, \ldots, x_r) = P\{X_1 \le x_1, \ldots, X_r \le x_r\}$. (Они могут понадобиться, например, в связи с тестом проверки серии, рассмотренным в следующем пункте.)
20. [BM50] Получите дальнейшую информацию об асимптотическом значении величины в левой части равенства (27) при больших n. [*Замечания.* Эмпирические данные показывают, что эта величина меньше e^{-2s^2} при любых n, но точно это неизвестно. Согласно работе Уиттла, она должна находиться где-то между $\exp(-2s^2 - (2s+s^3)/\sqrt{n})$ и $\exp(-2s^2 + 3.83s^3/\sqrt{n})$. Выражения, приведенные в табл. 2 для больших n, получены эмпирически на основе ограниченного набора данных; было бы полезно получить теоретически более надежные данные.]
21. [M40] Хотя, как было указано в тексте, критерий Колмогорова- Смирнова применяется в тех случаях, когда функция распределения $F(x)$ непрерывна, можно, конечно, попробовать вычислить K_n^+ и K_n^- для дискретных распределений. Проанализируйте вероятностное поведение K_n^+ и K_n^- для разных типов дискретных распределений $F(x)$. Сравните эффективность такого теста и критерия χ^2 в конкретных случаях.
22. [BM42] Обсудите использование "отношения наибольшего правдоподобия" в качестве альтернативы критерия χ^2 при проверке датчиков случайных чисел; проведите эксперименты, применяя эти два типа статистик.

3.3.2. Эмпирические тесты

В этом пункте будут описаны девять типов специализированных тестов, которые применяются для проверки случайности числовых последовательностей. Описание каждого теста состоит из двух частей: (a) краткой инструкции по применению теста и (b) изложения теории, лежащей в основе теста. (Читатели с недостаточной математической подготовкой могут при желании пропускать теоретические рассуждения. Наоборот, читателя со склонностью к математике могут заинтересовать сами по себе комбинаторные задачи, возникающие здесь, даже если он не собирается заниматься проверкой датчиков случайных чисел.)

Каждый тест применяется к последовательности случайных чисел

$$\langle U_n \rangle = U_0, U_1, U_2, \ldots, \tag{1}$$

которые должны быть распределены равномерно между нулём и единицей. Некоторые тесты предназначены в первую очередь для проверки целочисленных последовательностей, а не последовательностей действительных чисел типа (1). В этом случае проверке подвергается вспомогательная последовательность

$$\langle Y_n \rangle = Y_0, Y_1, Y_2, \ldots, \tag{2}$$

которая определяется следующим образом:

$$Y_n = \lfloor dU_n \rfloor. \tag{3}$$

Входящие в эту последовательность целые числа распределены равномерно между 0 и $d-1$, если числа **(1)** распределены равномерно между 0 и 1. Значение d может быть любым; например, на машине, работающей в двоичной системе, можно выбрать $d = 64 = 2^6$; тогда Y_n будут определяться шестью наиболее значимыми разрядами в двоичном представлении U_n. Чтобы тест был представительным, значение d должно быть достаточно большим, но не слишком, чтобы при реализации теста на машине не возникало затруднений.

Введенные здесь обозначения U_n, Y_n и d будут использоваться в данном разделе неоднократно. Значение d может быть разным в разных тестах.

A. Проверка равномерности (проверка частот). Первое требование, предъявляемое к последовательности **(1)**, заключается в том, чтобы числа U_n были действительно равномерно распределены между нулем и единицей. Проверить равномерность можно двумя способами: (a) с помощью критерия Колмогорова—Смирнова, взяв $F(x) = x$ при $0 \le x \le 1$; (b) выбрать какое-нибудь удобное значение d, например 100, на машине, работающей в десятичной системе, либо 64 и 128 на машине, работающей в двоичной системе, после чего использовать последовательность **(2)** вместо **(1)**. Затем для каждого целого r, $0 \le r < d$, подсчитать число значений $Y_j = r$ при $0 \le j < n$ и применить критерий χ^2 с $k = d$ и вероятностями $p_s = 1/d$.

Обоснование обоих подходов можно найти в предыдущем разделе.

B. Проверка серий. Проверяется равномерность и независимость пар следующих друг за другом случайных чисел. Для этого подсчитывается, сколько раз встречается каждая пара $(Y_{2j}, Y_{2j+1}) = (q, r)$ при $0 \le j < n$. Величины q и m могут принимать любые значения от 0 до d. Затем применяется критерий χ^2 с числом категорий $k = d^2$ и равными вероятностями $1/d^2$ во всех категориях. Значение d выбирается из тех же соображений, что и в предыдущем тесте, но в данном случае оно должно быть несколько меньше, так как применять критерий χ^2 следует при значениях n, существенно превышающих k (скажем, по крайней мере при $n > 5d^2$).

Очевидно, что можно обобщить этот тест на тройки, четверки и т. д. (см. упр. 2); однако значения d должны быть при этом резко уменьшены, чтобы число категорий не получалось слишком большим. Поэтому при объединении четырех и более элементов используются менее точные тесты, такие, как "наибольшее из t" или проверка комбинаций (они будут описаны ниже).

Отметим, что в этом тесте в n испытаниях используется $2n$ чисел последовательности **(2)**. Было бы ошибкой составлять в данном случае пары (Y_0, Y_1), (Y_1, Y_2), ..., (Y_{n-1}, Y_n); понятно ли читателю, почему? Для пар (Y_{2j+1}, Y_{2j+2}) можно было бы составить специальный тест и проверять каждую последовательность с помощью обоих тестов. С другой стороны, как показал Гуд, (*Annals af Mathematical Statistics,* **28**, **(1957), 262–264)**, если d—простое число, используются пары (Y_0, Y_1), (Y_1, Y_2), ..., (Y_{n-1}, Y_n) и с помощью обычного χ^2-метода вычисляются как статистика V_2, соответствующая проверке серий, *так и* статистика V_1 для проверки равномерности Y_0, Y_1, ..., Y_{n-1} с одним и тем же значением d, то при больших n разность $V_2 - 2V_1$ будет иметь распределение χ^2 с $(d-1)^2$ степенями свободы, хотя V_2 в данном случае имеет распределение, *отличное* от распределения χ^2 с $d^2 - 1$ степенями свободы.

C. Проверка интервалов. В этом тесте проверяется длина интервалов между появлениями значений U_j, принадлежащих некоторому заданному отрезку. Если α и β—два действительных числа, причем $0 \le \alpha < \beta \le 1$, то подсчитываются длины последовательностей U_j, U_{j+1}, ..., U_{j+r}, в которых только U_{j+r} лежит между α и β. (Такая последовательность из $r+1$ чисел определяет интервал длины r.)

Алгоритм G. *(Вычисление длин интервалов.)* Следующий алгоритм позволяет определить число интервалов длины 0, 1, ..., $t-1$ и полное число интервалов большей длины ($\ge t$) в последовательности **(1)**. Расчет продолжается до тех пор, пока не будет зарегистрировано всего n интервалов.

Picture: Рис 6. Реализация проверки интервалов (подобные алгоритмы применяются и при реализации тестов собирателя купонов и проверки на монотонность).

G1 [Начальная установка.] Установить $j \leftarrow -1$, $s \leftarrow 0$, а также COUNT$[r] \leftarrow 0$ для $0 \le r \le t$.

G2 [$r = 0$.] Установить $r \leftarrow 0$.

G3 [$\alpha \le U_j < \beta$?] Увеличить j на 1. Если $U_j \ge \alpha$ и $U_j < \beta$, перейти на **G5**.

G4 [Увеличить r.] Увеличить r на единицу, затем вернуться на **G3**.

G5 [Регистрация длины интервала.] (Обнаружен интервал длины r.) При $r \ge t$ прибавить 1 к COUNT$[t]$, в противном случае прибавить 1 к COUNT$[r]$.

G6 [Обнаружено n интервалов?] Прибавить 1 к s. Если $s < n$, вернуться на **G2**. ∎

После выполнения этого алгоритма $k = t + 1$ значений COUNT[0], COUNT[1], ..., COUNT[t] обрабатываются с помощью критерия χ^2 с использованием следующих значений вероятностей:

$$p_0 = p, \quad p_1 = p(1-p), \quad p_2 = p(1-p)^2, \quad \ldots, \quad p_{t-1} = p(1-p)^{t-1}, \quad p_t = p(1-p)^t. \tag{4}$$

Здесь $p = \beta - \alpha$ равно вероятности того, что $\alpha \le U_j < \beta$. Значения n и t обычно выбираются таким образом, чтобы средние значения COUNT[r] были не меньше 5.

Часто полагают либо $\alpha = 0$, либо $\beta = 1$, чтобы упростить шаг G3 описанного алгоритма. Частные случаи $(\alpha, \beta) = \left(0, \frac{1}{2}\right)$ или $\left(\frac{1}{2}, 1\right)$ иногда называют проверкой отклонений от среднего вниз или вверх.

Формулы (4) для вероятностей выводятся легко; мы предоставляем читателю возможность проделать это самостоятельно. Отметим, что в приведенном здесь алгоритме вычисляются длины заданного числа n *интервалов*; последовательность при этом может получиться произвольной длины. Можно, наоборот фиксировать длину проверяемой последовательности (см. упр. 5).

D. Покер-тест (проверка комбинаций). В "классическом" покер-тесте рассматриваются n групп из пяти следующих друг за другом целых чисел $(Y_{5j}, Y_{5j+1}, \ldots, Y_{5j+4})$, $0 \le j < n$. Выделяются следующие 7 типов комбинаций:

$abcde$ (все разные),
$aabcd$ (одна пара),
$aabbc$ (две пары),
$aaabc$ (три одного вида),
$aabb$ (полный сбор),
$aaaab$ (четыре одного вида),
$aaaaa$ (пять одного вида).

С помощью критерия χ^2 проверяется, соответствуют ли частоты комбинации вероятностям.

Для облегчения программирования желательно несколько упростить этот тест. Хорошим компромиссом может служить подсчет количества несовпадающих чисел в каждой пятерке Число категорий при этом равно пяти:

5 разных = все разные;
4 разных = одна пара;
3 разных = две пары, или три одного вида;
2 разных = полный сбор или четыре одного вида;
нет разных = пять одного вида.

Тест при этом почти не ухудшается, а алгоритм существенно упрощается.

В общем случае можно рассмотреть n групп по k следующих подряд чисел и подсчитывать число групп, в которых имеется r разных чисел. Затем применять критерий χ^2 с вероятностями

$$p_r = \frac{d(d-1)\ldots(d-r+1)}{d^k} \left\{ {k \atop r} \right\}. \tag{5}$$

(Числа Стирлинга $\left\{ {k \atop r} \right\}$ были определены в п. 1.2.6; там же приведены формулы, по которым их можно посчитать.) Так как при $r = 1$ или 2 вероятность p_r очень мала, эти категории можно объединить.

Приведенная выше формула для p_r получается следующим образом. Вычисляется, сколько групп по k чисел, каждое из которых может принимать значения от 0 до $d-1$, содержат r разных элементов. Затем результаты делятся на d^k—общее количество различных групп. Так как $d(d-1)\ldots(d-r+1)$ есть число размещений из d элементов по r, остается только показать, что $\left\{ {k \atop r} \right\}$—число способов, которыми можно распределить k элементов по r категориям. Окончательно формула (5) выводится с помощью результатов упр. 1.2.6-64.

E. Тест собирателя купонов. Этот тест и только что описанный связаны между собой примерно так же, как тесты проверки интервалов и проверки частот. Используется последовательность Y_0, Y_1, ... и определяются длины сегментов Y_{j+1}, Y_{j+2}, ..., Y_{j+r}, необходимых для того, чтобы собрать "полный набор" целых чисел от 0 до $d-1$. Это можно сделать с помощью следующего алгоритма.

Алгоритм C. *(Реализация теста собирателя купонов.)* Для заданной последовательности целых чисел $Y_0, Y_1, \ldots$ ($0 \le Y_j < d$) этот алгоритм определяет длины n последовательных сегментов по только

что сформулированному правилу. По окончании работы алгоритма в $\texttt{COUNT}[r]$ при $0 \le r < t$ находится число сегментов длины r, а в $\texttt{COUNT}[t]$—число сегментов длины $\ge t$.

C1 **[Начальная установка.]** Установить $j \leftarrow -1$, $s \leftarrow 0$, а также $\texttt{COUNT}[r] \leftarrow 0$ для $0 \le r \le t$.

C2 **[$q, r = 0$.]** Установить $q \leftarrow r \leftarrow 0$, а также $\texttt{OCCURS}[k] \leftarrow 0$ для $0 \le k < d$.

C3 **[Следующее испытание.]** Увеличить r и j на 1. Если $\texttt{OCCURS}[Y_j] \ne 0$, повторить эти операции.

C4 **[Полный набор?]** Установить $\texttt{OCCURS}[Y_j] \leftarrow 1$, $q \leftarrow q + 1$. (На данный момент обнаружено q разных значений. Если $q = d$, это значит, что получен полный набор.) При $q < d$ перейти на **C3**.

C5 **[Регистрация длины сегмента.]** При $r \ge t$ прибавить 1 к $\texttt{COUNT}[t]$, в противном случае прибавить 1 к $\texttt{COUNT}[r]$.

C6 **[Обнаружено n сегментов?]** Прибавить 1 к s. Если $s < n$, вернуться на **C2**. ∎

Пример использования этого алгоритма приведен в упр. 7. Можно представить себе, что речь идет о мальчике, собирающем d типов купонов. По одному купону содержится в каждой упаковке с кашей для завтрака, причем распределены они случайным образом. Мальчик ест кашу до тех пор, пока не соберет полный комплект купонов.

Содержимое счетчиков $\texttt{COUNT}[d]$, $\texttt{COUNT}[d+1]$, ..., $\texttt{COUNT}[t]$ обрабатывается с помощью критерия χ^2 с $k = t - d + 1$ степенями свободы после того, как алгоритм C завершил работу. Соответствующие вероятности равны:

$$p_r = \frac{d!}{d^r} \left\{ {r-1 \atop d-1} \right\}, \qquad d \le r < t, \qquad p_t = 1 - \frac{d!}{d^{t-1}} \left\{ {t-1 \atop d} \right\}. \tag{6}$$

Для вывода этих формул достаточно заметить, что, согласно (5), вероятность того, что сегмент длины r *неполный*, т. е. содержит не все из возможных d значений, равна

$$q_r = 1 - \frac{d!}{d^r} \left\{ {r \atop d} \right\}.$$

Тогда формулы (6) получаются непосредственно из соотношений $p_r = q_{r-1} - q_r$, $d \le r < t$ и $p_t = q_{t-1}$.

Формулы, возникающие в связи с различными *обобщениями* этого теста, рассмотрены в упр. 9 и 10, а также в работе Г. фон Шеллинга (*AMM*, **61** (1954), 306–311).

F. Проверка перестановок. Разделим исходную последовательность на n групп по t элементов в каждой: $(U_{jt}, U_{jt+1}, \ldots, U_{jt+t-1})$, $0 \le j < n$. В каждой группе возможно всего $t!$ вариантов относительного расположения чисел. Подсчитывается, сколько раз встречается каждое конкретное относительное расположение, после чего применяется критерий χ^2 с $k = t!$ и вероятностями $1/t!$ Например, при $t = 3$ число категорий равно шести: $U_{3j} < U_{3j+1} < U_{3j+2}$ или $U_{3j} < U_{3j+2} < U_{3j+1}$ или ... или $U_{3j+2} < U_{3j+1} < U_{3j}$. При этом предполагается, что точное равенство двух чисел невозможно: действительно, вероятность такого события равна нулю.

При реализации этого теста на машине удобно воспользоваться следующим алгоритмом, который интересен и сам по себе.

Алгоритм P. ***(Анализ перестановок.)*** Любому набору несовпадающих элементов $(U_1, \ldots, U_t)$ ставится в соответствие целочисленная функция $f(U_1, \ldots, U_t)$, такая, что

$$0 \le f(U_1, \ldots, U_t) < t!$$

и $f(U_1, \ldots, U_t) = f(V_1, \ldots, V_t)$ в том и только том случае, когда элементы в $(U_1, \ldots, U_t)$ и $(V_1, \ldots, V_t)$ расположены в одинаковом порядке. В программе используется вспомогательная таблица $C[1], \ldots, C[t]$.

P1 **[Начальная установка.]** Установить $r \leftarrow t$.

P2 **[Определить максимум.]** Определить максимальное из значений $\{ U_1, \ldots, U_r \}$. Допустим это U_s. Затем установить $C[r] \leftarrow s - 1$.

P3 **[Замена.]** Взаимозаменить $U_r \leftrightarrow U_s$.

P4 **[Уменьшение r.]** Вычесть из r единицу. Если $r > 0$, вернуться на **P2**.

P5 **[Вычисление f.]** Искомое значение функции определить по формуле

$$\begin{aligned} f &= C[t] + tC[t-1] + t(t-1)C[t-2] + \cdots + t!C[1] = \\ &= (\ldots((C[1] \times 2 + C[2]) \times 3 + C[3]) \times 4 + \cdots + C[t-1]) \times t + C[t]. \end{aligned} \tag{7}$$ ∎

∎

Алгоритм построен таким образом, что

$$0 \le C[r] < r \tag{8}$$

(см. шаг P2) и, в частности, $C[1]$ всегда равно нулю. Так как $C[r]$ может принимать r различных значений, общее количество разновидностей таблицы C равно $t!$. Из формул (7) и (8) легко видеть, что $0 \le f < t!$. К концу работы алгоритма величины $(U_1, \ldots, U_t)$ будут расставлены в порядке возрастания.

Для доказательства того, что результат f, полученный с помощью алгоритма P, характеризует единственным образом порядок следования чисел в исходной последовательности $(U_1, \ldots, U_t)$, покажем, что этот алгоритм можно использовать для решения обратной задачи. Пусть задано некоторое значение f и t элементов $(U_1, \ldots, U_t)$, причем $0 \le f < t!$ и $U_1 < U_2 < \ldots < U_t$. Установим $C[t] \leftarrow f \bmod t$, $C[t-1] \leftarrow \lfloor f/t \rfloor \bmod (t-1)$, $C[t-2] \leftarrow \lfloor \lfloor f/t \rfloor/(t-1) \rfloor \bmod (t-2)$ и т. д., выполняя, таким образом, шаг P5 в обратном порядке. Затем произведем замены $U_1 \leftrightarrow U_{C[1]+1}$, $U_2 \leftrightarrow U_{C[2]+1}$, $\ldots$, $U_t \leftrightarrow U_{C[f]+1}$ в указанном порядке. Легко видеть, что при этом ликвидируются последствия выполнения шагов P1–P4. Тот факт, что значение f позволяет восстановить исходный порядок последовательности, доказывает правильность алгоритма P.

G. Проверка на монотонность. Этот тест выявляет "неубывание" и "невозрастание", т. е. анализируются длины *монотонных* подпоследовательностей чисел в исходной последовательности.

В качестве примера рассмотрим последовательность из 10 цифр "1298536704"; отмечая вертикальными линиями начало и конец последовательности, а также все места, где $X_j > X_{j+1}$, получим

$$|1\,2\,9|8|5|3\,6\,7|0\,4|. \tag{9}$$

Здесь выделены все неубывающие подпоследовательности: первая длины 3, затем две по 1, еще одна длины 3 и затем 2. Алгоритм в упр. 12 показывает, как табулировать длины таких отрезков. В отличие от теста собирателя купонов или проверки интервалов (которые во многих отношениях похожи на этот тест) в данном случае *не следует применять критерий χ^2 для анализа полученных данных,* так как они *не являются* независимыми. После длинного отрезка чаще появляется короткий и наоборот. Из-за отсутствия независимости прямое применение критерия χ^2 становится незаконным. Вместо этого, после определения длин отрезков с помощью алгоритма, описанного в упр. 12, вычисляется статистика

$$V = \frac{1}{n} \sum_{1 \le i,j \le 6} (\texttt{COUNT}[i] - nb_i)(\texttt{COUNT}[j] - nb_j)a_{ij}, \tag{10}$$

где коэффициенты a_{ij} и b_i таковы:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} \\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} \end{pmatrix} = \begin{pmatrix} 4529.4 & 9044.9 & 13568 & 18091 & 22615 & 27892 \\ 9044.9 & 18097 & 27139 & 36187 & 45234 & 55789 \\ 13568 & 27139 & 40721 & 54281 & 67852 & 83685 \\ 18091 & 36187 & 54281 & 72414 & 90470 & 11580 \\ 22615 & 45234 & 67852 & 90470 & 113262 & 139476 \\ 27892 & 55789 & 83685 & 111580 & 139476 & 172860 \end{pmatrix}, \tag{11}$$

$$(\, b_1 \quad b_2 \quad b_3 \quad b_4 \quad b_5 \quad b_6 \,) = (\, \tfrac{1}{6} \quad \tfrac{5}{24} \quad \tfrac{11}{120} \quad \tfrac{19}{720} \quad \tfrac{29}{5040} \quad \tfrac{1}{840} \,).$$

(Здесь приведены приближенные значения коэффициентов; точные значения можно вычислить по приведенным ниже формулам.) *Статистика V в (10) должна иметь при больших n распределение χ^2 с шестью* (а не пятью) *степенями свободы.* Значение n должно быть равно, скажем, 4000 или больше. Аналогичный тест применяется для невозрастающих отрезков.

Довольно простой и более практичный способ проверки на монотонность приведен в упр. 14, но из чисто математических соображений полезнее будет разобрать этот весьма сложный тест.

Приступим к выводу формул для него. Пусть для данной перестановки из n элементов $Z_{pi} = 1$, если с позиции i начинается возрастающий отрезок с длиной не менее p, и $Z_{pi} = 0$ в противном случае. В качестве примера рассмотрим последовательность (9) с $n = 10$. В данном случае

$$Z_{11} = Z_{21} = Z_{31} = Z_{14} = Z_{15} = Z_{16} = Z_{26} = Z_{36} = Z_{19} = Z_{29} = 1,$$

а все остальные Z равны 0. Тогда

$$R'_p = Z_{p1} + Z_{p2} + \cdots + Z_{pn} \tag{12}$$

есть число отрезков длины не менее p, а

$$R_p = R'_p - R'_{p+1} \tag{13}$$

есть число отрезков, длина которых в точности равна p. Нам нужно вычислить среднее значение R_p, а также *ковариацию*[8]

$$\operatorname{covar}(R_p, R_q) = \operatorname{mean}((R_p - \operatorname{mean}(R_p))(R_q - \operatorname{mean}(R_q))),$$

которая служит мерой взаимозависимости R_p и R_q. Усреднение надо проводить по всем возможным $n!$ перестановкам.

Формулы (12) и (13) показывают, что искомые величины можно выразить через средние значения Z_{pi} и $Z_{pi}Z_{qj}$, поэтому в качестве первого шага запишем следующие соотношения (в предположении, что $i < j$):

$$\begin{aligned}
\frac{1}{n!}\sum Z_{pi} &= \begin{cases} (p+\delta_{i1})/(p+1)! & \text{при } i \le n-p+1; \\ 0 & \text{в противном случае;} \end{cases} \\
\frac{1}{n!}\sum Z_{pi}Z_{qj} &= \begin{cases} (p+\delta_{i1})q/(p+1)!(q+1)! & \text{при } i+p < j \le n-q+1; \\ (p+\delta_{i1})/(p+1)!q! - (p+q+\delta_{i1})/(p+q+1)! & \text{при } i+p = j \le n-q+1; \\ 0 & \text{в противном случае.} \end{cases}
\end{aligned} \tag{14}$$

Суммирование проводится здесь по всем возможным перестановкам. Чтобы показать, как выводятся эти формулы, рассмотрим самый сложный случай, когда $i+p = j \le n-q+1$ и $i > 1$. Отметим, что $Z_{pi}Z_{qj}$ равно либо 0, либо 1, так что суммирование сводится к тому, чтобы пересчитать все перестановки U_1, U_2, ..., U_n, в которых $Z_{pi} = Z_{qj} = 0$, а именно

$$U_{i-1} > U_i < \ldots < U_{i+p-1} > U_{i+p} < \ldots < U_{i+p+q-1}. \tag{15}$$

Число таких перестановок можно определить с помощью следующих соображений: существует $\binom{n}{p+1+q}$ способов выбора элементов последовательности, указанных в (15); имеется

$$(p+q+1)\binom{p+q}{p} - \binom{p+q+1}{p+1} - \binom{p+q+1}{1} + 1 \tag{16}$$

способов расположения их в порядке, указанном в (15), (см. упр. 13) и $(n-p-q-1)!$ способов расположения остальных элементов. Таким образом, надо умножить (16) на $\binom{n}{p+q+1} \times (n-p-q-1)!$ и после деления на $n!$ получается искомый результат.

Из соотношений (14) в результате довольно громоздких преобразований можно получить

$$\begin{aligned}
\operatorname{mean}(R'_p) &= \operatorname{mean}(Z_{p1} + \cdots + Z_{pn}) = \\
&= (n+1)p/(p+1)! - (p-1)/p!, \qquad 1 \le p \le n;
\end{aligned} \tag{17}$$

$$\begin{aligned}
\operatorname{covar}(R'_p R_q) &= \operatorname{mean}(R'_p R'_q) - \operatorname{mean}(R'_p)\operatorname{mean}(R'_q) = \\
&= \sum_{1 \le i,j \le n} \frac{1}{n!}\sum Z_{pi}Z_{pj} - \operatorname{mean}(R'_p)\operatorname{mean}(R'_q) = \\
&= \begin{cases} \operatorname{mean}(R'_t) + f(p,q,n) & \text{при } p+q \le n; \\ \operatorname{mean}(R'_t) - \operatorname{mean}(R'_p)\operatorname{mean}(R'_q) & \text{при } p+q > n; \end{cases}
\end{aligned} \tag{18}$$

где $t = \max(p,q)$, $s = p+q$ и

$$\begin{aligned}
f(p,q,n) = (n+1)&\left(\frac{s(1-pq)+pq}{(p+1)!(q+1)!} - \frac{2s}{(s+1)!}\right) + \\
&+ 2\left(\frac{s-1}{s!}\right) + \frac{(s^2-s-2)pq - s^2 - p^2q^2 + 1}{(p+1)!(q+1)!}.
\end{aligned} \tag{19}$$

[8] Мы оставляем обозначения автора для среднего, ковариации и т. п., следуя переводу 1-го тома (см. п. 1.2.10).—*Прим. перев.*

Конечно, пользоваться таким сложным выражением для ковариации очень неудобно, но другого выхода нет. С помощью этих формул легко вычислить

$$\begin{aligned}\operatorname{mean}(R_p) &= \operatorname{mean}(R'_p) - \operatorname{mean}(R'_{p+1}),\\ \operatorname{covar}(R_p, R'_q) &= \operatorname{covar}(R'_p, R'_q) - \operatorname{covar}(R'_{p+1}, R'_q),\\ \operatorname{covar}(R_p, R_q) &= \operatorname{covar}(R_p, R'_q) - \operatorname{covar}(R_p, R'_{q+1}).\end{aligned} \tag{20}$$

В работе Дж. Вольфовица (*Annals of Mathematical Statistics,* **15 (1944), 163–165)** доказано, что при $n \to \infty$ распределение величин R_1, R_2, …, R_{t-1}, R'_t стремится к нормальному со средним значением и ковариацией, приведенными выше. Отсюда следует, что можно пользоваться таким тестом. Для данной последовательности из n случайных чисел вычисляется R_p—число монотонных отрезков длины p, где $1 \le p < t$, а также R'_t—число отрезков длины $\ge t$. Введем обозначения

$$\begin{aligned}&Q_1 = R_1 - \operatorname{mean}(R_1), \ldots, Q_{t-1} = R_{t-1} - \operatorname{mean}(R_{t-1}),\\ &Q_t = R'_t - \operatorname{mean}(R'_t).\end{aligned} \tag{21}$$

Образуем матрицу ковариаций C, обозначая, например, $C_{13} = \operatorname{covar}(R_1, R_3)$, тогда как $C_{1t} = \operatorname{covar}(R_1, R'_t)$. При $t = 6$ получим

$$\begin{aligned}C = nC_1 + C_2 = \\ = n\begin{pmatrix} \frac{23}{180} & \frac{-7}{360} & \frac{-5}{336} & \frac{-433}{60480} & \frac{-13}{5670} & \frac{-121}{181440} \\ \frac{-7}{360} & \frac{2843}{20160} & \frac{-989}{20160} & \frac{-7159}{362880} & \frac{-10019}{1814400} & \frac{-1303}{907200} \\ \frac{-5}{336} & \frac{-989}{20160} & \frac{54563}{907200} & \frac{-21311}{1814400} & \frac{-62369}{19958400} & \frac{-7783}{9979200} \\ \frac{-433}{60480} & \frac{-7159}{362880} & \frac{-21311}{1814400} & \frac{886657}{39916800} & \frac{-257699}{239500800} & \frac{-62611}{239500800} \\ \frac{-13}{5670} & \frac{-10019}{1814400} & \frac{-62369}{19958400} & \frac{-257699}{239500800} & \frac{29874811}{5448643200} & \frac{-1407179}{21794572800} \\ \frac{-121}{181440} & \frac{-1303}{907200} & \frac{-7783}{9979200} & \frac{-62611}{239500800} & \frac{-1407179}{21794572800} & \frac{2134697}{1816214400} \end{pmatrix} + \\ + \begin{pmatrix} \frac{83}{180} & \frac{-29}{180} & \frac{-11}{210} & \frac{-41}{12096} & \frac{91}{25920} & \frac{41}{18144} \\ \frac{-29}{180} & \frac{-305}{4032} & \frac{319}{20160} & \frac{2557}{72576} & \frac{10177}{604800} & \frac{413}{64800} \\ \frac{-11}{210} & \frac{319}{20160} & \frac{-58747}{907200} & \frac{19703}{604800} & \frac{239471}{19958400} & \frac{39517}{9979200} \\ \frac{-41}{12096} & \frac{2557}{72576} & \frac{19703}{604800} & \frac{-220837}{4435200} & \frac{1196401}{239599800} & \frac{360989}{239500800} \\ \frac{91}{25920} & \frac{10177}{604800} & \frac{239471}{19958400} & \frac{1196401}{239500800} & \frac{-139126639}{7264857600} & \frac{4577641}{10897286400} \\ \frac{41}{18144} & \frac{413}{64800} & \frac{39517}{9979200} & \frac{360989}{239500800} & \frac{4577641}{10897286400} & \frac{-122953057}{21794572800} \end{pmatrix}.\end{aligned} \tag{22}$$

для $n \ge 14$. Затем найдем матрицу $A = (a_{ij})$, обратную матрице C, и вычислим $\sum_{1 \le i,j \le t} Q_i Q_j a_{ij}$. При больших n результат имеет распределение χ^2 с t степенями свободы.

Приведенная ранее матрица **(11)**—результат обращения матрицы C_1, представленной с пятью значащими цифрами. При больших n матрица A будет приближенно равна $(1/n)C_1^{-1}$. Делались попытки записать элементы матрицы, обратной к C_1 как рациональные числа, но это приводило к слишком большим величинам уже при $t = 4$. В частном случае при $n = 1000$ элементы матрицы **(11)** оказались примерно на 1% ниже точных значений, полученных в результате обращения матрицы **(22)**. Стандартный метод обращения матриц описан в п. **2.2.6**, упр. **18**.

H. Тест "наибольшее из t". Пусть $V_j = \max(U_{tj}, U_{tj+1}, \ldots, U_{tj+t-1})$ при $0 \le j < n$. Применим критерий Колмогорова-Смирнова к последовательности V_0, V_1, …, V_{n-1}, принимая в качестве теоретической функции распределения $F(x) = x^t$, **(**$0 \le x \le 1$**)**. Можно вместо этого проверять на равномерность последовательность V_0^t, V_1^t, …, V_{n-1}^t.

Для обоснования этого теста достаточно показать, что V_j распределены в соответствии с $F(x) = x^t$. Вероятность того, что $\max(U_1, U_2, \ldots, U_t) \le x$, равна вероятности того, что $U_1 \le x$ *и* $U_2 \le x$ *и* … *и* $U_t \le x$; следовательно, она равна произведению индивидуальных вероятностей $x \cdot x \cdot \ldots \cdot x = x^t$.

I. Последовательная корреляция. Вычислим статистику

$$C = \frac{n(U_0U_1 + \cdots + U_{n-2}U_{n-1} + U_{n-1}U_0) - (U_0 + \cdots + U_{n-1})^2}{n(U_0^2 + \cdots + U_{n-1}^2) - (U_0 + \cdots + U_{n-1})^2}. \tag{23}$$

Это "коэффициент последовательной корреляции", который служит мерой зависимости U_{j+1} от U_j. При очень больших n существует метод, позволяющий резко уменьшить время, которое тратится на расчет C **(**L. P. Schmid, *CACM,* **8 (1965), 115).**

Коэффициенты корреляции часто употребляются в статистике. Если имеется два набора величин U_0, U_1, …, U_{n-1} и V_0, V_1, …, V_{n-1}, то коэффициент корреляции между ними определяется следующим образом:

$$C = \frac{n\sum(U_jV_j) - (\sum U_j)(\sum V_j)}{\sqrt{(n\sum U_j^2 - (\sum U_j)^2)(n\sum V_j^2 - (\sum V_j)^2)}}. \tag{24}$$

Суммирование в этой формуле проводится по всем j в интервале $0 \le j < n$. Формула (23) получается в частном случае $V_j = U_{(j+1) \bmod n}$. [*Замечание.* Знаменатель выражения (24) равен нулю при $U_0 = U_1 = \cdots = U_{n-1}$ или $V_0 = V_1 = \cdots = V_{n-1}$; мы исключаем этот случай из рассмотрения.)

Коэффициент корреляции всегда лежит между -1 и $+1$. Когда он равен нулю или очень мал, это служит указанием на независимость величин U_j и V_j, а когда он равен ± 1, эти величины связаны друг с другом линейной зависимостью, т. е. для любого j справедливо равенство $V_j = m \pm aU_j$ при некоторых постоянных a и m (см. упр. 17).

Таким образом, желательно, чтобы значение C, определенное формулой (23), было близко к нулю. В действительности, из-за того, что U_0U_1 и U_1U_2, конечно, не являются независимыми, коэффициент последовательной корреляции не должен быть в *точности* равен нулю (см. упр. 18). "Хорошим" можно считать значение C, лежащее между $\mu_n - 2\sigma_n$ и $\mu_n + 2\sigma_n$, где

$$\mu_n = \frac{-1}{(n-1)}, \quad \sigma_n = \frac{1}{n-1}\sqrt{\frac{n(n-3)}{n+1}}, \qquad n > 2. \tag{25}$$

Значение C должно находиться в этих пределах в 95% всех случаев.

Формулы (25) пока имеют предположительный характер, так как точное распределение C в том случае, когда U распределены равномерно, неизвестно. Случай нормального распределения величин U рассмотрен в работе У. Диксона (*Annals of Mathematical Statistics,* **15** (1944), 119–144). Опыт показывает, что использование формул для математического ожидания и среднеквадратичного отклонения, соответствующих нормальному распределению, т. е. формул (25), не приводит к большой ошибке. Известно, что $\lim_{n\to\infty} \sqrt{n}\sigma_n = 1$; см. также статью Андерсона и Уокера (*Annals of Mathematical Statistics,* **35** (1964), 1296–1303), в которой получены более общие результаты о последовательной корреляции в *зависимых* последовательностях.

J. Проверка подпоследовательностей. Нередко внешняя программа устроена так, что ей требуется каждый раз некоторое определенное количество случайных чисел. Например, если в программе есть три случайные величины X, Y и Z, для определения их значений каждый раз нужно будет генерировать три случайных числа. Для таких приложений важно, чтобы случайной была любая последовательность, полученная в результате выбора каждого *третьего* числа исходной последовательности. Если программа запрашивает каждый раз q чисел, то с помощью тестов, описанных выше, можно проверять не исходную последовательность $U_0, U_1, U_2, \ldots$, а в отдельности каждую из подпоследовательностей

$$U_0, U_q, U_{2q}, \ldots; \quad U_1, U_{q+1}, U_{2q+1}, \ldots; \quad \ldots; \quad U_{q-1}, U_{2q-1}, \ldots . \tag{26}$$

Опыт показывает, что при использовании линейного конгруэнтного метода характеристики таких подпоследовательностей практически никогда не бывают хуже, чем у исходной последовательности, кроме случаев, когда q и число, используемое в качестве модуля, содержат достаточно большой общий множитель. Так, на машинах с двоичной арифметикой при использовании значений m, равных размеру слова, из всех $q < 16$ самые плохие результаты получаются при $q = 8$; на машинах с десятичной арифметикой можно ожидать неудовлетворительных результатов при $q = 10$. (Это можно объяснить отчасти, исходя из понятия мощности последовательности, так как такие значения q будут, вообще говоря, понижать ее мощность.)

K. Замечания исторического характера и дальнейшее обсуждение. Статистические критерии возникали естественным образом в процессе научной работы, когда появлялась необходимость "принять" или "отвергнуть" какую-либо гипотезу, касающуюся экспериментальных данных. Лучшими среди работ, посвященных проверке на случайность искусственных последовательностей чисел, являются две статьи М. Кендалла и Б. Бэбингтон-Смита [*Journal of the Royal Statistical Society,* **101** (1938), 147–166, и приложение к этому журналу, **6** (1939), 51–61]. В этих работах описана проверка случайных цифр от 0 до 9, а не действительных случайных чисел; для этой цели предложены проверка частот, серий, интервалов, а также покер-тест (хотя проверка серий производится неверно). Кендалл и Бэбингтон-Смит использовали также разновидность теста собирателя купонов, но в том виде, который описан в данной книге, этот тест был предложен Гринвудом в 1955 г.

Тест проверки монотонности имеет довольно интересную историю. Первоначально в нем регистрировались одновременно длины отрезков как с возрастающими, так и убывающими числами (эти отрезки чередуются). Следует отметить, что этот тест, так же как проверка перестановок, не требует, чтобы значения U были распределены равномерно; требуется только, чтобы вероятность того, что $U_i = U_j$, равнялась нулю при $i \ne j$, так что эти тесты можно применять к многим типам случайных последовательностей. В примитивной форме этот тест впервые был предложен в работе

[J. Bienaynie, *Comptes Rendus,* **81** (Paris: Acad. Sciences, 1875), 417–423]. Около 60 лет спустя Кермэк и Мак-Кендрик написали две большие работы, посвященные этому вопросу [*Proc. Royal Society Edinburgh,* **57** (1937), 228–240, 332–376]. В качестве примера в них показано с помощью проверки на монотонность, что колебания количества осадков в Эдинбурге в период с 1785 г. по 1930 г. носили случайный характер (хотя они исследовали только средние и стандартные отклонения отрезков монотонности). С тех пор этот тест периодически использовался на практике, но только в 1944 г. было показано, что применять его в сочетании с критерием χ^2 нельзя. В работе Левена и Вольфовица [*Annals of Mathematical Statistics,* **15** (1944), 58–69] была приведена правильная формулировка теста (с чередующимися отрезками возрастания и убывания) и указана ошибочность его первоначальной формулировки. Вариант теста, изложенный в данной книге, когда анализируются длины отрезков или только с возрастанием, или только с убыванием, наиболее удобен для реализации на вычислительной машине, поэтому формулы для других вариантов не приводятся (см. обзор Barton D. E., Mallows C. L., *Annals of Math. Statistics,* **36** (1965), 236–260].

Из всех описанных здесь тестов проверка частот и проверка последовательной корреляции— самые слабые, в том смысле, что при испытании с помощью этих тестов почти все датчики случайных чисел дают удовлетворительные результаты. Вкратце теоретическое обоснование этого будет дано в § 3.5 (см. упр. 3.5-26). К сравнительно сильным тестам относится проверка на монотонность: результаты упр. 3.3.3-23, 24 показывают, что при недостаточно больших значениях m последовательности, полученные с помощью линейного конгруэнтного метода, имеют повышенную длину отрезков монотонности, так что этот тест определенно полезен.

Вероятно, у читателя возникает вопрос: *"Зачем так много тестов?"*. Может создаться впечатление, что на испытания датчиков случайных чисел тратится больше машинного времени, чем на выработку случайных чисел в процессе решения прикладных задач! Это неверно, хотя случаи чрезмерного увлечения проверками возможны.

Необходимость достаточно разнообразного набора тестов многократно отмечалась в литературе. В частности, указывалось, что последовательности, полученные с помощью некоторых разновидностей метода середины квадрата, хорошо проходят проверку частот, интервалов, комбинаций, но оказываются совершенно негодными при проверке серий. Известно, что датчики, основанные на линейном конгруэнтном методе, удовлетворяют при малых значениях m многим тестам, но не удовлетворяют проверке на монотонность, так как дают слишком мало отрезков единичной длины. Тест "наибольшее из t" также позволяет выявить плохие датчики, которые со всех других точек зрения ведут себя вполне приемлемо.

Вероятно, основная причина, по которой необходима всесторонняя проверка датчиков, заключается в следующем. Если кто-то пользуется чужим датчиком случайных чисел, то при любом недоразумении он будет винить этот датчик, а не свою программу. Нужно, чтобы автор датчика мог *доказать,* что случайные числа удовлетворяют всем требованиям. С другой стороны, если вы пишете датчик для себя, а не для общего пользования, можно не тратить сил на его проверку; во всяком случае, если за основу взять какой-либо из алгоритмов, рекомендованных в этой главе, с большой вероятностью этот датчик будет вполне удовлетворительным.

Упражнения

1. [10] Почему при проверке серий (см. п. B) следует использовать пары (Y_0, Y_1) (Y_2, Y_3), ..., (Y_{2n-2}, Y_{2n-1}), а не (Y_0, Y_1), (Y_1, Y_2), ..., (Y_{n-1}, Y_n)?
2. [10] Покажите, как обобщить проверку серий с пар на тройки, четверки и т. д.

>3. [M20] Сколько в среднем потребуется перебрать значений U при проверке интервалов (алгоритм G), прежде чем будет обнаружено n интервалов, в предположении, что последовательность действительно случайная? Каково стандартное отклонение этой величины?

4. [12] Покажите, что при проверке интервалов законно пользоваться вероятностями (4).
5. [M23] При "классической" проверке интервалов, описанной Кендаллом и Бэбингтон-Смитом, N значений U, подлежащих проверке, используются для построения циклической последовательности, в которой U_{N+j} совпадает с U_j. Если n чисел из U_0, ..., U_{N-1} попадают в интервал $\alpha \le U_j < \beta$, то в циклической последовательности имеется n интервалов. Пусть Z_r—число интервалов длины r, если $0 \le r < t$, а Z_t—число интервалов длины $\ge t$. Покажите, что величина $V = \sum_{0\le r\le t}(Z_r - np_r)^2 \Big/ np_r$ должна иметь при $N \to \infty$ распределение χ^2 с t степенями свободы, где p_r определены выражением (4).
6. [40](X. Гирингер) Анализ частот появления различных цифр в десятичном представлении числа $e = 2.71828\ldots$ дает для первых 2000 цифр значение $\chi^2 = 1.06$. Это указывает на то, что частоты лежат значительно ближе к средним значениям, равным 200, чем можно было бы ожидать при случайном распределении (вероятность значений $\chi^2 \ge 1.15$ равна 99.9%). Если применить тот же

тест к первым 10 000 цифрам, получается разумное значение $\chi^2 = 8.61$; но факт столь правильного распределения первых 2000 цифр остается удивительным. Произойдет ли то же самое при представлении e в системе счисления с другим основанием? (См. *AMM,* **72** (1965), 483–500.)

7. [08] Проверьтеспомощью теста собирателя купонов (алгоритм C) при $d = 3$ и $n = 7$ следующую последовательность: 1101221022120202001212201010201121. Каковы длины семи подпоследовательностей?

>8. [M22] Сколько в среднем надо перебрать значений U в тесте собирателя купонов (алгоритм C) для получения n полных наборов, если последовательность случайная? Определите также стандартное отклонение этой величины. (*Указание:* см. формулу 1.2.9-28.)

9. [M21] Обобщите тест собирателя купонов так, чтобы дальнейший поиск прекращался после того, как найдено w различных значений, где $w \le d$—положительное целое число. Какие вероятности надо использовать в этом случае вместо (6).

10. [M23] Выполните упр. 8 для обобщенного теста собирателя купонов,

11. [00] В представлении (9) для конкретной последовательности отмечены возрастающие отрезки; каковы убывающие отрезки для той же последовательности?

12. [22] Задано n различных чисел $U_0, U_1, \ldots, U_{n-1}$. Запишите алгоритм, который определяет длины всех отрезков возрастания в этой последовательности. В результате работы алгоритма в COUNT$[r]$ при $1 \le r \le 5$ должно находиться число отрезков длины r, а в COUNT$[6]$—число отрезков длины 6 или более.

13. [M23] Покажите, что (16) есть число перестановок типа (15) из $p + q + 1$ различных элементов.

14. [M15] Если мы будем "выбрасывать" элемент, непосредственно следующий за отрезком монотонности, так что в случае $X_j > X_{j+1}$ очередной отрезок монотонности начнется с X_{j+2}, то длины таких отрезков будут независимыми и можно будет воспользоваться обычным критерием χ^2 (вместо весьма сложного метода, приведенного в тексте). Какими должны быть соответствующие вероятности длин отрезков монотонности для этого упрощенного теста?

15. [M20] Почему значения V_0^t, V_1^t, $\ldots$, V_{n-1}^t в тесте "наибольшее из t" должны быть распределены равномерно между нулем и единицей?

>16. [15] (a) Пусть требуется проделать вычисления для теста "наибольшее из t" при разных значениях t. Обозначим $Z_{jt} = \max(U_j, U_{j+1}, \ldots, U_{j+t-1})$. Студент Смышлёный обнаружил остроумный способ перехода от последовательности $Z_{0(t-1)}$, $Z_{1(t-1)}$, $\ldots$ к последовательности Z_{0t}, Z_{1t}, $\ldots$, требующий минимальных вычислений. Попробуйте найти этот способ.

(b) Он же решил изменить метод "наибольшее из t" так, чтобы $V_j = \max(U_j, \ldots, U_{j+t-1})$; другими словами, $V_j = Z_{jt}$, а не $V_j = Z_{(tj)t}$, как указано в тексте. При этом он рассуждал так: *все* должны иметь одинаковое распределение, поэтому тест должен стать только сильнее, если использовать все Z_{jt}, $0 \le j < n$, а не одно из каждых t значений. Однако при проверке равномерности значений V_j^t получились крайне высокие значения статистики V, причем при увеличении t они становились еще больше. Как это могло получиться?

17. [M25](a) Заданы произвольные числа $X_1, \ldots, X_n, Y_1, \ldots, Y_n$, причем

$$\bar{x} = \frac{1}{n}\sum_{1\le k\le n} X_k, \qquad \bar{y} = \frac{1}{n}\sum_{1\le k\le n} Y_k.$$

Обозначим $X'_k = X_k - \bar{x}$, $Y'_k = Y_k - \bar{y}$. Покажите, что коэффициент корреляции C, вычисленный по формуле (24), равен

$$\sum_{1\le k\le n} X'_k Y'_k \Big/ \sqrt{\sum_{1\le k\le n} {X'_k}^2} \sqrt{\sum_{1\le k\le n} {Y'_k}^2}.$$

(b) Пусть $C = N/D$, где N и D—числитель и знаменатель только что приведенного выражения. Покажите, что $N^2 \le D^2$, т. е. $-1 \le C \le +1$; получите формулу для разности $D^2 - N^2$. (*Указание:* см. упр. 1.2.3-30.)

(c) Покажите, что при $C = \pm 1$ для $1 \le k \le n$ можно найти такие постоянные a, b, m не все равные нулю, что $aX_k + bY_k = m$.

18. [M20] (a) Покажите, что при $n = 2$ коэффициент последовательной корреляции (23) *всегда* равен -1 (если знаменатель не равен нулю). (b) Покажите, что при $n = 3$ коэффициент последовательной корреляции всегда равен $-\frac{1}{2}$. (c) Покажите, что знаменатель в (23) равен нулю тогда и только тогда, когда $U_0 = U_1 = \cdots = U_{n-1}$.

19. [M40] Чему равны среднее значение и стандартное отклонение коэффициента последовательной корреляции (23), когда $n = 4$, а значения U независимы и распределены равномерно между нулем и единицей?

20. **[M50]** Определите распределение коэффициента последовательной корреляции (23) при произвольном n, предполагая, что случайные величины U_j независимы и распределены равномерно между нулем и единицей.
21. **[19]** Какое значение f даст алгоритм P для перестановки $(1, 2, 9, 8, 5, 3, 6, 7, 0, 4)$?
22. **[18]** Для какой перестановки набора целых чисел $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ алгоритм P даст значение $f = 1024$?

3.3.3. *Теоретические тесты

Хотя любой датчик случайных чисел можно проверить с помощью методов, описанных в предыдущем пункте, гораздо лучше иметь "*априорные* тесты", т. е. теоретические результаты, которые позволяют сказать заранее, каким будет исход испытаний датчика. Такого рода теоретические результаты позволяют значительно глубже понять способы порождения случайных чисел, чем эмпирические данные, полученные методом "проб и ошибок". Здесь будут более детально изучены линейные конгруэнтные последовательности; зная, какими должны быть случайные числа еще до их выработки, мы имеем больше надежды правильно выбрать значения a, m и c.

Построить теорию линейного конгруэнтного метода очень трудно, хотя определенных успехов в этой области достичь удалось. Полученные до сих пор результаты относятся главным образом к *статистическому анализу последовательности на полном периоде*. В такой постановке задача имеет смысл не для всех тестов; например, при проверке равномерности распределения на всем периоде будут получаться слишком хорошие результаты. Однако проверку серий, интервалов, перестановок и т. д. можно успешно проводить и на полном периоде.

Для начала докажем простое *априорное* утверждение, касающееся наименее сложного варианта проверки перестановок Суть нашей первой теоремы состоит в том, что если датчик обладает высокой мощностью, то примерно в половине случаев будет выполняться неравенство $X_{n+1} < X_n$.

Теорема P. *Пусть X_0, a, c и m определяют линейную конгруэнтную последовательность с максимальным периодом; пусть $b = a - 1$, а d—наибольший общий делитель m и b. Тогда вероятность того, что $X_{n+1} < X_n$, равна $\frac{1}{2} + r$, где*

$$r = (2(c \bmod d) - d)/2m;$$

следовательно, $|r| < d/2m$.

Техника, которая используется при доказательстве этой теоремы, интересна и сама по себе. Начнем с того, что введем

$$s(x) = (ax + c) \bmod m. \tag{1}$$

Тогда $X_{n+1} = s(X_n)$ и требуется подсчитать число таких целых x, что $0 \le x < m$ и $s(x) < x$ (так как каждое такое число встречается, где-то в пределах периода). Мы хотим показать, что это число равно

$$\frac{1}{2}(m + 2(c \bmod d) - d). \tag{2}$$

При $a = 1$ теорема справедлива: легко видеть, что в этом случае $s(x) < x$ только при $x + c \ge m$, а это имеет место c раз.

Лемма A. *Пусть α и β—действительные числа, а n—целое число. В этом случае*

$$\alpha < n \le \beta \qquad \textit{тогда и только тогда, когда } \lfloor\alpha\rfloor < n \le \lfloor\beta\rfloor; \tag{3}$$

$$\alpha \le n < \beta \qquad \textit{тогда и только тогда, когда } \lceil\alpha\rceil \le n < \lceil\beta\rceil. \tag{4}$$

Эти формулы следуют непосредственно из определений "потолка" и "пола"; см. упр. 1.2.4-3.∎

Для любого целого x, $0 \le x < m$, определим $k(x) = \lfloor(ax + c)/m\rfloor$; тогда $s(x) = ax + c - k(x)m$ и $k(x) \le (ax + c)/m$, откуда следует, что неравенство $s(x) < x$ эквивалентно

$$(k(x)m - c)/a \le x < (k(x)m - c)/b. \tag{5}$$

Далее, эти неравенства единственным образом определяют $k(x)$, поскольку из них вытекают неравенства

$$\frac{ax + c}{m} \le k(x) < \frac{bx + c}{m} < \frac{ax + c}{m} + 1.$$

В соответствии с леммой A можно утверждать, что $s(x) < x$ и $0 \le x < m$ тогда и только тогда, когда выполнено одно из взаимоисключающих условий:

$$\textbf{(i)}\ \lceil(km - c)/a\rceil \le x < \lceil(km - c)/b\rceil \qquad \text{для какого-либо } k,\ 0 < k < a; \tag{6}$$

$$\textbf{(ii)}\ \lceil m - c/a\rceil \le x < m.$$

Число таких целых x равно

$$m + \sum_{0<k\le b} \lceil (km-c)/b \rceil - \sum_{0<k\le a} \lceil (km-c)/a \rceil. \tag{7}$$

Теорема будет доказана, если мы покажем, что (7) равно (2). Суммы в выражении (7) можно вычислить так, как описано в упр. 1.2.4-37; мы воспользуемся другим методом, чтобы продемонстрировать еще один способ решения задач такого типа.

Определим прежде всего следующие функции:

$$\delta(z) = \lfloor z \rfloor + 1 - \lceil z \rceil = \begin{cases} 1, & \text{если } z \text{ целое;} \\ 0, & \text{если } z \text{ не целое;} \end{cases} \tag{8}$$

$$((z)) = z - \lfloor z \rfloor - \frac{1}{2} + \frac{1}{2}\delta(z) = z - \lceil z \rceil + \frac{1}{2} - \frac{1}{2}\delta(z). \tag{9}$$

Последняя представляет собой "пилообразную функцию", используемую в теории рядов Фурье; ее график показан на рис. 7. Преимущества работы с функцией $((z))$ по сравнению с $\lfloor z \rfloor$ или $\lceil z \rceil$ определяются тем, что она обладает рядом полезных

Picture: Рис. 7. Пилообразная функция $((z))$.

свойств:

$$((-z)) = -((z)); \tag{10}$$

$$((z+n)) = ((z))\, z, \qquad \text{если } n \text{ целое;} \tag{11}$$

$$((z)) + \left(\left(z + \frac{1}{n}\right)\right) + \cdots + \left(\left(z + \frac{n-1}{n}\right)\right) = ((nz)), \qquad \text{если } n\text{—положительное целое число} \tag{12}$$

(см. упр. 2). Меняя обозначения, можно преобразовать (7) следующим образом:

$$\frac{1}{2}\left(m - 1 + \sum_{0<k\le a} \delta\left(\frac{km-c}{a}\right) - \sum_{0<k\le b} \delta\left(\frac{km-c}{b}\right)\right) + \sum_{0<k\le a} \left(\left(\frac{km-c}{a}\right)\right) - \sum_{0<k\le b} \left(\left(\frac{km-c}{b}\right)\right). \tag{13}$$

На вид эта формула хуже той, с которой мы начинали, но в действительности все входящие в нее суммы легко вычисляются. Так как m и a взаимно просты, мы знаем, что $(km-c) \bmod a$ принимает в определенном порядке каждое из значений $0, 1, \ldots, a-1$ при $0 < k \le a$, так что

$$\sum_{0<k\le a} \delta\left(\frac{km-c}{a}\right) = 1, \qquad \sum_{0<k\le a} \left(\left(\frac{km-c}{a}\right)\right) = 0.$$

(Справедливость второй формулы вытекает из равенств (10) и (11).) Далее, если числа m и b—не взаимно простые, а c и b—взаимно простые, то $(km-c)/b$ не может быть целым, поэтому вторая сумма в (13) исчезает. Наконец, имеют место равенства

$$\begin{aligned}\left(\left(\frac{km-c}{b}\right)\right) &= \left(\left(\frac{km - d\lfloor c/d \rfloor - c \bmod d}{b}\right)\right) = \\ &= \left(\left(\frac{km/d - \lfloor c/d \rfloor}{b/d}\right)\right) - \frac{c \bmod d}{b} + \frac{1}{2}\delta\left(\frac{km/d - \lfloor c/d \rfloor}{b/d}\right),\end{aligned}$$

так как $0 \le (c \bmod d)/b < 1/(b/d)$. Сумма

$$\sum_{0<k\le b} \left(\left(\frac{km-c}{b}\right)\right)$$

равна, следовательно, $\frac{1}{2}d - c \bmod d$, так как m/d и b/d взаимно просты. Теорема Р доказана. ∎

Доказательство теоремы Р демонстрирует трудность задачи *априорной* проверки случайных чисел, однако оно в то же время показывает и ее выполнимость. Далее, из этой теоремы следует что практически при любом выборе a и c неравенство $X_{n+1} < X_n$ будет выполняться с нужной частотой, во

всяком случае на всем периоде, кроме тех случаев, когда d велико. Большие значения d соответствуют малой мощности, что, как мы знаем, нежелательно.

Следующая теорема дает более важный критерий для выбора a и c; в ней рассматривается *коэффициент последовательной корреляции* на полном периоде. Этот коэффициент был определен в п. 3.3.2 (см. формулу (23)):

$$C = \left(m \sum_{0\le x<m} xs(x) - \left(\sum_{0\le x<m} x \right)^2 \right) \Big/ \left(m \sum_{0\le x<m} x^2 - \left(\sum_{0\le x<m} x \right)^2 \right). \tag{14}$$

Пусть x'—такой элемент, что $s'(x) = 0$. Тогда

$$s(x) = m\left(\left(\frac{ax+c}{m}\right)\right) + \frac{m}{2}, \qquad \text{если } x \ne x'. \tag{15}$$

Для упрощения записи формул воспользуемся одной важной функцией, которая часто возникает во многих математических задачах:

$$\sigma(h,k,c) = 12 \sum_{0\le j<k} \left(\left(\frac{j}{k}\right)\right)\left(\left(\frac{hj+c}{k}\right)\right); \tag{16}$$

она называется *обобщенной суммой Дедекинда.*

Так как

$$\sum_{0\le x<m} x = \frac{m(m-1)}{2} \quad \text{и} \quad \sum_{0\le x<m} x^2 = \frac{m(m-1)(2m-1)}{6},$$

то формулу (14) довольно просто преобразовать к виду

$$C = \frac{m\sigma(a,m,c) - 3 + 6(m-x'-c)}{m^2-1}. \tag{17}$$

Поскольку m обычно очень велико, можно отбросить члены порядка $1/m$ и записать приближенное равенство

$$C \approx \sigma(a,m,c)/m \tag{18}$$

с ошибкой, меньшей $6/m$ по абсолютной величине.

Проверка последовательной корреляции сводится при этом к определению суммы Дедекинда $\sigma(a,m,c)$. Вычисление $\sigma(a,m,c)$ непосредственно по формуле (16) едва ли проще прямого вычисления коэффициента корреляции, но, к счастью, существуют простые методы быстрого получения значений сумм Дедекинда.

Лемма B. *("Закон взаимности" для сумм Дедекинда.) Если* $0 \le c < h$, $0 \le c < k$ *и числа* h *и* k *взаимно простые, то*

$$\sigma(h,k,c) + \sigma(k,h,c) = \frac{h}{k} + \frac{k}{h} + \frac{1}{hk} + \frac{6c^2}{hk} - 3. \tag{19}$$

Доказательство. Мы предоставляем читателю показать, что в этих предположениях справедливо

$$\sigma(h,k,c) + \sigma(k,h,c) = \sigma(h,k,0) + \sigma(k,h,0) + 6c^2/hk \tag{20}$$

(см. упр. 6). Теперь достаточно рассмотреть только случай $c = 0$.

Приведем принадлежащее Л. Карлицу доказательство, основанное на использовании комплексных корней из единицы. Существует более простое доказательство, требующее только элементарных манипуляций с суммами (см. упр. 7); однако приведенный ниже способ более поучителен, так как в нем используется техника, которая может быть полезна и в других задачах такого типа.

Определим полиномы $f(x)$ и $g(x)$ следующим образом:

$$\begin{aligned} f(x) &= 1 + x + \cdots + x^{k-1} = (x^k - 1)/(x-1), \\ g(x) &= x + 2x^2 + \cdots + (k-1)x^{k-1} = xf'(x) = \\ &= kx^k/(x-1) - x(x^k-1)/(x-1)^2. \end{aligned} \tag{21}$$

Если $\omega = e^{2\pi i/k}$—комплексный корень степени k из единицы то, согласно (1.2.9-13),

$$\frac{1}{k}\sum_{0<j\le k}\omega^{jr}g(\omega^j x) = rx^r \qquad \text{при } 0 \le r < k. \tag{52}$$

Положим $x = 1$; тогда $g(\omega^j x) = k/(\omega^j - 1)$ при $j \ne k$ и $k(k-1)/2$ при $j = k$. Отсюда следует, что

$$r \bmod k \sum_{0<j<k}\frac{\omega^{-jr}}{\omega^j - 1} + \frac{1}{2}(k-1), \qquad \text{если } r\text{—целое.} \tag{23}$$

(Так как, согласно (22), правая часть равна r при $0 \le r < k$ и при добавлении к r кратного k она не изменяется.) Следовательно,

$$\left(\left(\frac{r}{k}\right)\right) = \frac{1}{k}\sum_{0<j<k}\frac{\omega^{-jr}}{\omega^j - 1} - \frac{1}{2k} + \frac{1}{2}\delta\left(\frac{r}{k}\right). \tag{24}$$

Эта важная формула, которая справедлива для любых целых r, позволяет свести $((r/k))$ к сумме, включающей корни степени k из единицы; при этом возникает целый ряд новых возможностей. В частности, имеет место следующая формула:

$$\sigma(h,k,0) + \frac{3(k-1)}{k^2} = \frac{12}{k^2}\sum_{0<r<k}\sum_{0<i<k}\sum_{0<j<k}\frac{\omega^{-ir}}{\omega^i - 1}\frac{\omega^{-jh_r}}{\omega^j - 1}. \tag{25}$$

Правую часть этой формулы можно упростить, проведя суммирование по r; если $s \bmod k \ne 0$, то $\sum_{0\le r<k}\omega^{rs} = f(\omega^s) = 0$. Формула (25) сводится к

$$\sigma(h,k,0) + \frac{3(k-1)}{k} = \frac{12}{k}\sum_{0<j<k}\frac{1}{(\omega^{-j^h} - 1)(\omega^j - 1)}. \tag{26}$$

Аналогичная формула с заменой ω на $\zeta = e^{2\pi i/h}$ справедлива для $\sigma(k,h,0)$.

Вопрос о том, что делать дальше с суммой в (26), не очевиден; однако существует элегантный способ дальнейших преобразований, основанный на том факте, что каждый член суммы является функцией ω^j, $0 < j < k$, и по существу суммируются все корни степени k из единицы, кроме 1. Заметим, что для любых несовпадающих комплексных чисел $x_1, x_2, \ldots, x_n$ справедливо следующее равенство:

$$\sum_{1<j\le n}\frac{1}{(x_j - x_1)\ldots(x_j - x_{j-1})(x - x_j)(x_j - x_{j+1})\ldots(x_j - x_n)} = \frac{1}{(x - x_1)\ldots(x - x_n)}, \tag{27}$$

которое следует из обычного разложения правой части на элементарные дроби. Далее, если $q(x) = (x - y_1)(x - y_2)\ldots(x - y_m)$, то

$$q'(y_j) = (y_j - y_1)\ldots(y_j - y_{j-1})(y_j - y_{j+1})\ldots(y_j - y_m). \tag{28}$$

Это равенство часто используется для упрощения выражений такого типа, как левая часть (27). Если h и k взаимно просты, то все значения $\omega, \omega^2, \ldots, \omega^{k-1}, \zeta, \zeta^2, \ldots, \zeta^{h-1}$ не совпадают друг с другом. Следовательно, можно рассматривать (27) как частный случай полинома $(x - \omega)\ldots(x - \omega^{k-1})(x - \zeta)\ldots(x - \zeta^{h-1}) = (x^k - 1)(x^h - 1)/(x - 1)^2$, откуда следует равенство

$$\frac{1}{h}\sum_{0<j<h}\frac{\zeta^j(\zeta^j - 1)^2}{(\zeta^{ik} - 1)(x - \zeta^j)} + \frac{1}{k}\sum_{0<j<k}\frac{\omega^j(\omega^j - 1)^2}{(\omega^{jh} - 1)(x - \omega^j)} = \frac{(x-1)^2}{(x^h - 1)(x^k - 1)}. \tag{29}$$

Из этого равенства вытекает много интересных следствий, и оно приводит к целому ряду соотношений взаимности для сумм, таких, как (26). Например, если продифференцировать (29) дважды по x и положить $x = 1$, то получится

$$\frac{2}{h}\sum_{0<j<h}\frac{\zeta^j(\zeta^j - 1)^2}{(\zeta^{jk} - 1)(1 - \zeta^j)^3} + \frac{2}{k}\sum_{0<j<k}\frac{\omega^j(\omega^j - 1)^2}{(\omega^{jk} - 1)(1 - \omega^j)^3} = \frac{1}{6}\left(\frac{h}{k} + \frac{k}{h} + \frac{1}{hk}\right) + \frac{1}{2} - \frac{1}{2h} - \frac{1}{2k}.$$

Заменяя в этих суммах j на $h-j$ и $k-j$, получим [ср. (26)]

$$\frac{1}{6}\left(\sigma(k,h,0)+\frac{3(h-1)}{h}\right)+\frac{1}{6}\left(\sigma(h,k,0)+\frac{3(k-1)}{k}\right)=\frac{1}{6}\left(\frac{h}{k}+\frac{k}{h}+\frac{1}{hk}\right)+\frac{1}{2}-\frac{1}{2h}-\frac{1}{2k},$$

что эквивалентно требуемому результату. ∎

Лемма C. *Если числа h и k взаимно простые и $0<c<k$, то*

$$\sigma(h,k,c)=\sigma(h,k,c \bmod h)+\frac{6r}{k}(c+c \bmod h-k)+d, \tag{30}$$

где $r=\lfloor c/h\rfloor$, $d=0$ при $c \bmod h\neq 0$ и $d=3$ при $c \bmod h=0$.

Доказательство. При любом c имеем

$$\begin{aligned}\sigma(h,k,c+h)&=12\sum_{0\le j<k}\left(\left(\frac{j}{k}\right)\right)\left(\left(\frac{hj+c+h}{k}\right)\right)=\\&=12\sum_{0\le j<k}\left(\left(\frac{j-1}{k}\right)\right)\left(\left(\frac{hj+c}{k}\right)\right)=\\&=\sigma(h,k,c)+6\left(\left(\frac{h+c}{k}\right)\right)+6\left(\left(\frac{c}{k}\right)\right),\end{aligned} \tag{31}$$

так как, согласно (9),

$$\left(\left(\frac{j-1}{k}\right)\right)=\left(\left(\frac{j}{k}\right)\right)-\frac{1}{k}+\frac{1}{2}\sigma\left(\frac{j+1}{k}\right)+\frac{1}{2}\sigma\left(\frac{j}{k}\right). \tag{32}$$

Пусть теперь $c_0=c \bmod h$, так что $0<c=c_0+rh<k$. Тогда

$$\begin{aligned}\sigma(h,k,c)&=\sigma(h,k,c_0)+12\sum_{0\le j<r}\left(\left(\frac{c_0+hj}{k}\right)\right)+6\left(\left(\frac{c}{k}\right)\right)-6\left(\left(\frac{c_0}{k}\right)\right)=\\&=\sigma(h,k,c_0)+12\sum_{0\le j<r}\left(\frac{c_0+hj}{k}-\frac{1}{2}\right)+6\left(\frac{c}{k}-\frac{1}{2}\right)-6\left(\frac{c_0}{k}-\frac{1}{2}\right)+d.\end{aligned}$$

Простое суммирование завершает доказательство. ∎

Леммы B и C позволяют сформулировать следующую эффективную процедуру вычисления $\sigma(h,k,c)$ в случае, когда h и k взаимно просты.

Шаг 1. Добавлять (или вычитать) кратные k к (из) h и c (если это необходимо), с тем чтобы привести их к интервалам $-\frac{k}{2}<h\le\frac{k}{2}$, $-\frac{k}{2}<c\le\frac{k}{2}$ (или к интервалам $0<h\le k$, $0\le c<k$). Такая операция законна, потому что по определению

$$\sigma(h\pm nk,k,c)=\sigma(h,k,c)=\sigma(h,k,c\pm nk). \tag{33}$$

Шаг 2. Если h или c отрицательны, сделать их положительными, используя равенства

$$\begin{aligned}\sigma(-h,k,c)&=-\sigma(h,k,c),\\\sigma(h,k,-c)&=\sigma(h,k,c).\end{aligned} \tag{34}$$

Шаг 3. Если $k=1$ или $k=2$, то $\sigma(h,k,c)=0$ (так как $((j/1))=((j/2))=0$).

Шаг 4. Если $c>h$, заменить c на $c \bmod h$ с помощью соотношения (30) леммы C.

Шаг 5. Теперь соблюдены условия леммы B, так что имеем

$$\sigma(h,k,c)=-3+\frac{h}{k}+\frac{k}{h}+\frac{1+6c^2}{hk}-\sigma(k,h,c). \tag{35}$$

Для определения $\sigma(k,h,c)$ вернуться к шагу 1.

Число необходимых итераций обычно невелико; последовательные значения h и k ведут себя так же, как последовательность значений, получаемая при определении с помощью алгоритма Евклида (см. п. 4.5.2) наибольшего общего делителя h и k. Рассмотрим несколько примеров.

Пример 1. *Найти коэффициент последовательной корреляции для случая* $m = 2^{35}$, $a = 2^{34} + 1$, $c = 1$.

Решение. Согласно (17), имеем

$$C = (2^{35}\sigma(2^{34}+1, 2^{35}, 1) - 3 + 6(2^{35} - (2^{34}-1) - 1))/(2^{70}-1). \tag{36}$$

Выполняя шаги 1 и 2, получаем

$$(\sigma(2^{34}+1, 2^{35}, 1) = -\sigma(2^{34}-1, 2^{35}, 1).$$

Согласно шагу 5:

$$\sigma(2^{34}-1, 2^{35}, 1) = -3 + (2^{34}-1)/2^{35} + 2^{35}/(2^{34}-1) + 7/2^{35}(2^{34}-1) - \sigma(2^{35}, 2^{34}-1, 1).$$

Согласно шагу 1:

$$\sigma(2^{35}, 2^{34}-1, 1) = \sigma(2, 2^{34}-1, 1).$$

Теперь шаг 5 дает

$$\sigma(2, 2^{34}-1, 1) = -3 + 2/(2^{34}-1) + (2^{34}-1)/2 + 7/2(2^{34}-1) - \sigma(2^{34}-1, 2, 1)$$

и

$$\sigma(2^{34}-1, 2, 1) = 0.$$

В результате получаем, что

$$C = \frac{1}{4} + \varepsilon, \qquad |\varepsilon| < 2^{-67}. \tag{37}$$

Такая корреляция, безусловно, неприемлема. Конечно, этот датчик обладает слишком малой мощностью; мы уже отвергли его ранее как неслучайный.

Пример 2. *Определить приблизительно коэффициент последовательной корреляции при* $m = 10^{10}$, $a = 10001$, $b = 2113248658$.

Решение. Так как $C \approx \sigma(a, m, c)/m$, делаем следующие вычисления:

$$\begin{aligned}
\sigma(10001, 10^{10}, 2113248653) &= \sigma(10001, 10^{10}, 7350) - 6(211303)(7886743997)/10^{10};\\
\sigma(10001, 10^{10}, 7350) &\approx -3 + 10^{10}/10001 - \sigma(10^{10}, 10001, 7350);\\
\sigma(10^{10}, 10001, 7350) &= \sigma(100, 10001, 7350) =\\
&= \sigma(100, 10001, 50) - 6(73)(2601)/10001;\\
\sigma(100, 10001, 50) &\approx -3 + 10001/100 + 100/10001 - \sigma(10001, 100, 50);\\
\sigma(10001, 100, 50) &= \sigma(1, 100, 50) = -50,02.\\
C &\approx (-3 + 999900,01 - 97,02 - 50,02 + 113,91 - 99895,60)/10^{10} =\\
&= -0,000000003172.
\end{aligned} \tag{38}$$

Такое значение C, конечно, удовлетворяет любым требованиям. Но мощность этого датчика равна всего 3, *так что, несмотря на отсутствие последовательной корреляции, его нельзя считать хорошим источником случайных чисел.* Отсутствие корреляции— необходимое, но не достаточное условие!

Пример 3. *Оценить последовательную корреляцию при любых* a, m, c.

Первую фазу приведенных выше расчетов можно проделать в общем виде. Пусть $c_0 = c \bmod a$.

$$\begin{aligned}
\sigma(a, m, c) &= \sigma(a, m, c_0) + \frac{6(c-c_0)}{am}(c + c_0 - m) =\\
&= -3 + \frac{a}{m} + \frac{m}{a} + \frac{1}{am} + \frac{6c^2}{am} - \frac{6(c-c_0)}{a} - \sigma(m, a, c_0).
\end{aligned} \tag{39}$$

Согласно упр. 12, $|\sigma(m, a, c_0)| < a$, поэтому

$$C \approx \frac{\sigma(a, m, c)}{m} \approx \frac{1}{a}\left(1 - 6\frac{c}{m} + 6\left(\frac{c}{m}\right)^2\right). \tag{40}$$

При этом мы пренебрегли членами

$$\frac{3}{m}\left(1 - 2\frac{c_0}{a} + \frac{a - x' - c}{m}\right) + \frac{1}{m}\sigma(m, a, c_0) + O\left(\frac{1}{m^2}\right),$$

так что можно сказать, что погрешность выражения (40) не превышает a/m. Главный вклад в ошибку дает член $\sigma(m, a, c_0)/m$. Полученные результаты можно суммировать следующим образом.

Теорема S. *Коэффициент последовательной корреляции для линейной конгруэнтной последовательности с максимальным периодом определяется приближенным выражением (40) с ошибкой, меньшей чем* $(a+6)/m$. *Точное значение можно эффективно вычислить с помощью выражения (17), используя леммы B и C.* ∎

Из соотношения (40) вытекают некоторые заслуживающие упоминания следствия. Во-первых, оно доказывает, что надо избегать малых значений a. С другой стороны, большие значения a еще не гарантируют того, что корреляция будет мала, как было показано в примере 1; ошибка выражения (40) может достигать a/m, и тогда при больших значениях a/m приближение становится плохим. Если $a \approx \sqrt{m}$, значения коэффициента последовательной корреляции ограничены величиной $2/\sqrt{m}$.

Соотношение (40) помогает и при выборе значения c. До сих пор мы знали относительно c только одно: числа c и m должны быть, взаимно простыми. Если, кроме того,

$$\frac{c}{m} \approx \frac{1}{2} - \frac{1}{6}\sqrt{3} \approx 0.21132\,48654\,051871 \approx \approx (0.15414\ 54272\ 33746\ 34354\ 55716)_8, \tag{41}$$

коэффициент последовательной корреляции будет небольшим, так как корни уравнения $1 - 6x + 6x^2 = 0$ равны $\frac{1}{2} \pm \frac{1}{6}\sqrt{3}$. Этим критерием можно пользоваться, если отсутствуют другие соображения относительно выбора c.

До сих пор речь шла о корреляции между X_n и X_{n+1}. Желательно также иметь низкую корреляцию между X_n и X_{n+2}, и вообще было бы неплохо, если бы корреляция между X_n и X_{n+t} была низкой, скажем, для $1 \le t \le 10$. Ранее было показано (см. формулу (3.2.1-6)), что

$$X_{n+t} = (a_t X_n + c_t) \bmod m, \tag{42}$$

где

$$a_t = a^t \bmod m, \qquad c_t = (a^t - 1)c/(a-1) \bmod m. \tag{43}$$

С помощью подведенных выше формул можно вычислить коэффициент корреляции между X_n *и* X_{n+t}, *если вместо* a *и* c *подставить* a_t *и* c_t. Конечно, c_t уже не будет удовлетворять условию (41), но с этим придется смириться.

Приближенное выражение (40) впервые получил Р. Ковэю (*JACM,* **7** (1960), 72–74) в результате усреднения по всем *действительным числам* x между 0 и m, а не только по целым значениям (см. упр. 21). Методы, позволяющие получить точный результат, были позднее развиты М. Гринбергером (*Math. Comp.,* **15** (1961), 383–389) и Б. Янссоном (*BIT,* **4** (1964), 6–27). В их формулах суммы Дедекинда не использовались. Янссон составил таблицы для коэффициента последовательной корреляции, но, к сожалению, он рассматривал слишком простые множители, пользоваться которыми не рекомендуется. Он показал например, что коэффициент корреляции между X_n и X_{n+t} меньше 0.000003 для датчика с $m = 2^{35}$, $a = 2^{24} + 5$, $c = 1$ при всех $t \le 2500$. С помощью спектрального теста (см. п. 3.3.4) можно показать, что пользоваться этим датчиком не следует; тем не менее результат Янссона—хорошее свидетельство того, что у выбранного наугад датчика, основанного на линейном конгруэнтном методе и обладающего высокой мощностью, можно ожидать низкую последовательную корреляцию. [*Замечание.* Янссон также получил формулы для коэффициента последовательной корреляции в последовательностях с $c = 0$ и множителем, обеспечивающим максимальный период при $m = 2^e$. Эти результаты в сущности аналогичны рассмотренным в этой книге; см. упр. 3.2.1 2-9.]

Суммы Дедекинда $\sigma(h, k, c)$ и закон взаимности (для частного случая $c = 0$) были впервые рассмотрены Р. Дедекиндом в 1892 г. в его работе, посвященной эллиптическим функциям. Эта функция использовалась многими авторами; список литературы по данному вопросу можно найти в работах У. Дитера [*Journal für die reine und angewandte Mathematik,* **201** (1959), 37–70], а также Г. Радемахера и Э. Уайтмэна [*American Journal of Mathematics,* **63** (1941), 377–407].

Еще несколько *априорных* тестов будет описано в упражнениях к этому разделу. Главный вывод, который следует из них, заключается в следующем: *множитель в линейном конгруэнтном методе должен быть достаточно велик.* См. также упр. 3.3.4-7, где приводится дальнейшее развитие теоремы Р.

Упражнения

(ПЕРВАЯ ЧАСТЬ)

1. [M07] Объясните, почему выражение (7) равно числу значений x, для которых $s(x) < x$, $0 \le x < m$.

2. **[M24]** Докажите справедливость соотношения (12)
3. **[BM22]** Разложите в ряд Фурье (выразите через синусы и косинусы) функцию $f(x) = ((x))$.
>4. **[M18]** Каково наибольшее возможное значение d (в обозначениях теоремы P), если $m = 10^{10}$, а мощность датчика равна 10?
5. **[M21]** Выведите формулу (17).
6. **[M27]** Пусть $hh' + kk' = 1$ (a) Покажите, не пользуясь леммами B и C, что

$$\sigma(h,k,c) = \sigma(h,k,0) + 12\sum_{0\le j<c}\left(\left(\frac{h'j}{k}\right)\right) + 6\left(\left(\frac{h'c}{k}\right)\right)$$

для всех $c \ge 0$. (b) Покажите, что при $0 \le c < h$ и $0 \le c < k$

$$\left(\left(\frac{h'c}{k}\right)\right) + \left(\left(\frac{k'c}{h}\right)\right) = \frac{c}{hk}.$$

(c) В предположениях леммы B покажите справедливость равенства (20).

>7. **[M24]** Предложите доказательство закона взаимности (19) при $c = 0$, основанное на идеях упр. 1.2.4-45.
>8. **[M34]** Пусть

$$\rho(p,q,r) = 12\sum_{0\le j<r}\left(\left(\frac{jp}{r}\right)\right)\left(\left(\frac{jq}{r}\right)\right).$$

Обобщая метод, использованный при доказательстве леммы B, докажите следующее красивое соотношение (полученное Радемахером): "Если p, q и r попарно взаимно просты, то

$$\rho(p,q,r) + \rho(q,r,p) + \rho(r,p,q) = \frac{p}{qr} + \frac{q}{rp} + \frac{r}{pq} = 3."$$

(Закон взаимности для сумм Дедекинда при $c = 0$ является частным случаем этого равенства при $r = 1$.)

9. **[M40]** Существует ли простое доказательство соотношения Радемахера (см. упр. 8), которое является частным случаем доказательства, рассмотренного в упр. 7?
10. **[M20]** Докажите соотношения (34) об изменении знака параметров, входящих в $\sigma(h,k,c)$.
11. **[M30]** Формулы в тексте дают возможность вычислить $\sigma(h,k,c)$, когда h и k взаимно просты. В общем случае положим $hh' \equiv d \pmod k$, где d—наибольший общий делитель h и k. Покажите, что

$$\sigma(h,k,c) = \sigma(h/d, k/d, \lfloor c/d \rfloor) + \delta,$$

где $\delta = 0$ при $c \bmod d = 0$ и

$$\delta = 6\left(\left(\frac{\lfloor c/d\rfloor h'd}{k}\right)\right)$$

при $c \bmod d \ne 0$.

12. **[M24]** Покажите, что если h и k взаимно просты, то $|\sigma(h,k,c)| < (k-1)(k-2)/k$.
13. **[M22]** В (37) приведено *приближенное* значение для коэффициента последовательной корреляции при $m = 2^{35}$, $a = 2^{34} + 1$, $c = 1$. Чему равно точное значение?
>14. **[M24]** При проверке последовательной корреляции у линейного конгруэнтного датчика с $m = 2^{35}$, $a = 2^{18} + 1$, $c = 1$ были исследованы три отрезка последовательности по 1000 чисел в каждом; во всех случаях коэффициент корреляции получился довольно большим: от 0.2 до 0.3. Каким будет коэффициент последовательной корреляции для этого датчика по всему периоду в 2^{35} чисел?
15. **[M22]** Определите точное значение коэффициента последовательной корреляции в частном случае $a = 1$.
16. **[M24]** Как мы уже знаем, линейный, конгруэнтный метод дает плохие случайные числа при $a = 1$. Покажите с помощью теоремы P, что, несмотря на это, можно подобрать такое c, что вероятность выполнения неравенства $(X_{n+1} < X_n)$ будет очень близка к $1/2$ даже при $a = 1$. Можно также подобрать такое c, что коэффициент последовательной корреляция будет очень низким. Можно ли выбрать значение c так, чтобы выполнялись оба эти условия одновременно?
17. **[M21]** Объясните, как выбрать такое значение a, чтобы как c, так и c_2 удовлетворяли приближенному соотношению (41), где c_2—приращение последовательности, заданное (43).
>18. **[M35]** Положим

$$S(h,k,c,z) = \sum_{0\le j<z}\left(\left(\frac{hj+c}{k}\right)\right).$$

При взаимно простых h и k получите "формулы взаимности", которые помогут эффективно вычислять значения этой функции, подобно тому, как с помощью лемм B и C вычисляются значения $\sigma(h,k,c)$. [*Указание:* см. упр. 6.]

>19. [M35] *Проверку серий,* описанную в предыдущем раздето можно проводить на полном периоде. Пусть α, β, γ, δ—целые числа, причем $0 \le \alpha < \beta \le m$, $0 \le \gamma < \delta \le m$. Выведите соотношение для вероятности того, что $\alpha \le X_n < \beta$ и $\gamma \le X_{n+1} < \delta$.

20. [M24] Обобщите (26) таким образом, чтобы получалось выражение для $\sigma(h,k,c)$.

Упражнения

(ВТОРАЯ ЧАСТЬ) Во многих случаях точные расчеты с целыми числами слишком сложны, однако можно попытаться при определении вероятностей проводить усреднение не по целым значениям, а по всему интервалу изменения x. Хотя такие результаты будут приближенными, они прольют свет на обсуждаемый предмет.

Удобнее всего иметь дело с числами U_n, заключенными между нулем и единицей; для линейных конгруэнтных последовательностей $U_n = X_n/m$, так что $U_{n+1} = \{aU_n + \theta\}$, где $\theta = c/m$, а $\{x\}$ означает $x \bmod 1$. Например, формула для коэффициента последовательной корреляции принимает вид

$$C = \left(\int_0^1 x\{ax+\theta\}\,dx - \left(\int_0^1 x\,dx\right)^2\right) \Big/ \left(\int_0^1 x^2\,dx - \left(\int_0^1 x\,dx\right)^2\right).$$

>1. [BM23] (Р. Ковэю.) Чему равно C в только что приведенной формуле?

>2. [M22] Пусть a—целое число и $0 \le \theta < 1$. Если x—действительное число, находящееся между 0 и 1, а $s(x) = \{ax+\theta\}$, то какова вероятность, что $s(x) < x$? (Это аналог теоремы P для действительных чисел.)

3. [M28] Предыдущее упражнение дает вероятность того, что $U_{n+1} < U_n$. Какова вероятность того, что $U_{n+2} < U_{n+1} < U_n$, если U_n—случайные действительные числа, заключенные между нулем и единицей?

4. [M29] Сохраняя условия предыдущей задачи и положив $\theta = 0$, покажите, что вероятность выполнения неравенств $U_n > U_{n+1} > \cdots > U_{n+t-1}$ в точности равна

$$\frac{1}{t!}\left(1+\frac{1}{a}\right)\ldots\left(1+\frac{t-2}{a}\right).$$

Какова средняя длина убывающего ряда чисел, начинающегося со случайно выбранного между нулем и единицей числа U_n?

>5. [M25] Пусть α, β, γ, δ—действительные числа, причем $0 \le \alpha < \beta \le 1$, $0 \le \gamma < \delta \le 1$. Какова вероятность того, что $\alpha \le x < \beta$ и $\gamma \le s(x) < \delta$, если выполнены предположения упр. 22? (Это аналог упр. 19 для действительных чисел.)

6. [M21] Рассмотрим датчик, основанный на методе Фибоначчи с $U_{n+1} = \{U_n + U_{n-1}\}$. Предполагая, что U_1 и U_2 выбираются независимо случайным образом между нулем и единицей, найдите вероятность того, что $U_1 < U_2 < U_3$, $U_1 < U_3 < U_2$, $U_2 < U_1 < U_3$ и т. д. [*Указание:* разделить "единичный квадрат", т. е. точки на плоскости $\{\,(x,y) \mid 0 \le x,y < 1\,\}$, на шесть частей, в зависимости от соотношения между x, y и $\{x+y\}$, и определить площадь каждой части.]

7. [M27] Пусть исходные числа для датчика, описанного в предыдущем упражнении, выбираются независимо в единичном квадрате, после чего большее из них берется в качестве U_0, а меньшее—U_1. Определите вероятность того, что U_1 окажется начальной точкой возрастающего ряда длины k, т. е. $U_0 > U_1 < \ldots < U_k > U_{k+1}$. Сравните результат с соответствующими вероятностями для случайной последовательности.

>8. [M35] Для линейного конгруэнтного датчика мощности 2 выполняется условие $X_{n-1} - 2X_n + X_{n+1} \equiv (a-1)c \pmod m$ (см. формулу (3.2.1.3-5)). Рассмотрите датчик, который является непрерывным аналогом, положив $U_{n+1} = \{\alpha + 2U_n - U_{n-1}\}$. Так же как в упр. 26, разделите единичный квадрат на части, доказывающие возможные соотношения между U_{n-1}, U_n и U_{n+1} для каждой пары (U_{n-1}, U_n). Существует ли значение α, при котором каждое из шести возможных соотношений между этими числами осуществляется с вероятностью $1/6$, если U_{n-1} и U_n выбираются случайно в единичном квадрате?

3.3.4. Спектральный тест

Важный тест для проверки случайности полученных на ЭВМ числовых последовательностей сформулировали в 1965 г. Р. Ковэю и Р. Макферсон. Этот тест замечателен тем, что все известные плохие датчики, основанные на линейном конгруэнтном методе, были им *забракованы,* в то время,

как все хорошие датчики прошли удовлетворительно! Безусловно, это наиболее совершенный из имеющихся тестов, в связи с чем он заслуживает особого внимания.

Спектральный тест обладает свойствами как "эмпирических", так и "теоретических" тестов, рассмотренных в предыдущих разделах. Как и в теоретических тестах, в нем рассматриваются величины, усредненные по всему периоду. С другой стороны, для получения результатов испытаний требуются машинные расчеты, что делает его похожим на эмпирические тесты.

Обоснование этого теста требует использования математики в значительных дозах. Читателю без особой склонности к математике рекомендуется перейти прямо к подпункту D данного пункта, где приводится описание вполне конкретного спектрального теста.

A. Теория, лежащая в основе теста. Математическим обоснованием спектрального теста служит "конечное преобразование Фурье" функции, определенной на конечном множестве. Одномерный случай конечного преобразования Фурье уже использовался в предыдущем разделе при доказательстве леммы 3.3.3 B; рассмотрим теперь технику преобразования Фурье в общем случае.

Для любой принимающей комплексные значения функции $F(t_1, t_2, \ldots, t_n)$, определенной для всех комбинаций целых чисел t_k, где $0 \le t_k < m$ при $1 \le k \le n$, введем *преобразование Фурье*

$$f(s_1, s_2, \ldots, s_n) = \sum_{0 \le t_1, \ldots, t_n < m} \exp\left(\frac{-2\pi i}{m}(s_1 t_1 + \cdots + s_n t_n)\right) F(t_1, \ldots, t_n). \tag{1}$$

Функция f определена для всех комбинаций целых s_k; это периодическая функция в том смысле, что $f(s_1, \ldots, s_n) = f(s_1 \bmod m, \ldots, s_n \bmod m)$. Чтобы связать это определение с формулами предыдущего раздела, отметим, что

$$\exp\left(\frac{-2\pi i}{m}(s_1 t_1 + s_2 t_2 + \cdots + s_n t_n)\right) = \omega^{-(s_1 t_1 + s_2 t_2 + \cdots + s_n t_n)},$$

если $\omega = e^{2\pi i/m}$—корень m-й степени из единицы. Термин "преобразование" в данном случае оправдан, так как исходную функцию $F(t_1, \ldots, t_n)$ можно восстановить по ее преобразованию $f(s_1, \ldots, s_n)$ и представить в следующем виде (см. упр. 1):

$$F(t_1, \ldots, t_n) = \frac{1}{m^n} \sum_{0 \le s_1, \ldots, s_n < m} \exp\left(\frac{2\pi i}{m}(s_1 t_1 + \cdots + s_n t_n)\right) f(s_1, \ldots, s_n). \tag{2}$$

Эту формулу можно записать, используя синусы и косинусы для большего сходства с бесконечными рядами Фурье. Величина $(1/m^n) f(s_1, \ldots, s_n)$ является амплитудой n-мерной комплексной плоской волны с частотами $s_1/m, \ldots, s_n/m$, если $F(t_1, \ldots, t_n)$ представлена в виде (2).

Из соотношений (1) и (2) следует, что теоретически можно определить любые свойства F по ее преобразованию f, и наоборот. Часто бывает удобнее работать с преобразованием Фурье функции, а затем произвести обратное преобразование, чтобы вывести неочевидные свойства исходной функции.

Попробуем применить эту концепцию к проблеме выработки случайных чисел. Пусть задана бесконечная последовательность целых чисел X_0, X_1, X_2, ..., причем $0 \le X_k < m$, и пусть n—фиксированное (обычно небольшое) положительное целое число. Определим

$$F(t_1, \ldots, t_n) = \lim_{N \to \infty} \frac{1}{N} \sum_{0 \le k < N} \delta_{X_k t_1} \delta_{X_{k+1} t_2} \ldots \delta_{X_{k+n-1} t_n}. \tag{3}$$

Эта функция равна предельной плотности числа появлений комбинации $(t_1, \ldots, t_n)$ в виде n следующих друг за другом элементов последовательности $X_0, X_1, X_2, \ldots$. Так как нас интересуют периодические последовательности $X_0, X_1, X_2, \ldots$, можно считать, что предел в (3) существует, причем для вычисления предела можно взять N равным длине периода. В действительно случайной последовательности с равномерным распределением любые комбинации из t чисел должны появляться с одинаковой частотой, так что $F(t_1, \ldots, t_n)$ должно быть равно $1/m^n$ для любых $t_1, \ldots, t_n$.

Производя преобразование Фурье функции (3), получим

$$f(s_1, \ldots, s_n) = \lim_{N \to \infty} \frac{1}{N} \sum_{0 \le k < N} \exp\left(\frac{-2\pi i}{m}(s_1 X_k + s_2 X_{k+1} + \cdots + s_n X_{k+n-1})\right). \tag{4}$$

Когда мы имеем дело с действительно случайной последовательностью, должен получиться образ константы $1/m^n$; так что *для случайной последовательности мы должны получить*

$$f(s_1, \ldots, s_n) = \begin{cases} 1, & \text{если } s_1 \equiv \cdots \equiv s_n \equiv 0 \pmod m; \\ 0 & \text{в противном случае.} \end{cases} \tag{5}$$

В теоретических тестах, рассмотренных ранее, используются средние значения по полному периоду. Для определения среднего от какой-либо функции, зависящей от n следующих друг за другом элементов, достаточно в принципе знать $F(t_1, \ldots, t_n)$. Например, вероятность того, что $X_k < X_{k+1}$, равна величине $F(t_1, t_2)$, просуммированной по всем $0 \le t_1 < t_2 < m$; точно так же случай $n = 2$ позволяет определить коэффициент последовательной корреляции (см. упр. 5).

Так как образ $f(s_1, \ldots, s_n)$ содержит в себе всю информацию об $F(t_1, \ldots, t_n)$, его также можно использовать в *любом* теоретическом тесте. Следовательно, можно ожидать, что величину отклонения $f(s_1, \ldots, s_n)$ от значений (5), отвечающих действительно случайной последовательности, можно использовать для проверки случайности.

Для линейных конгруэнтных последовательностей $f(s_1, \ldots, s_n)$ имеет очень простой вид, чего нельзя сказать об $F(t_1, \ldots, t_n)$. Рассмотрим линейную конгруэнтную последовательность с параметрами a, m, c и X_0, имеющую *максимальную длину периода* в соответствии с теоремой 3.2.1.2A. Для такой последовательности

$$f(s_1, \ldots, s_n) = \frac{1}{m} \sum_{0 \le k < m} \exp\left(\frac{-2\pi i}{m}(s_1 X_k + s_2 X_{k+1} + \cdots + s_n X_{k+n-1})\right) = \\ = \frac{1}{m} \sum_{0 \le k < m} \exp\left(\frac{-2\pi i}{m}\left(s(a) X_k + \frac{s(a) - s(1)}{a-1} c\right)\right), \tag{6}$$

где

$$s(a) = s_1 + s_2 a + s_3 a^2 + \cdots + s_n a^{n-1}, \tag{7}$$

так как

$$X_{k+r} \equiv a^r X_k + \frac{a^r - 1}{a-1} c \pmod m$$

согласно формуле (3.2.1-6). Мы предположили, что последовательность имеет максимальный период, так что в ней встречаются все значения X_k; следовательно, (6) сводится к

$$\frac{1}{m} \sum_{0 \le k < m} \exp\left(\frac{-2\pi i}{m}\left(s(a) k + \frac{s(a) - s(1)}{a-1} c\right)\right).$$

Это сумма геометрической прогрессии, поэтому можно записать следующую основную формулу:

$$f(s_1, \ldots, s_n) = \exp\left(\frac{-2\pi i c}{m}\left(\frac{s(a) - s(1)}{a-1}\right)\right) \delta\left(\frac{s(a)}{m}\right), \tag{8}$$

где $\delta(x) = 1$, если x—целое, и $\delta(x) = 0$ в противном случае.

Напомним, что формула (2) позволяет интерпретировать $f(s_1, s_2, \ldots, s_n)/m^n$ физически как амплитуду n-мерной комплексной плоской волны

$$\omega(t_1, \ldots, t_n) = \exp\left(2\pi i \left(\frac{s_1}{m} t_1 + \cdots + \frac{s_n}{m} t_n\right)\right). \tag{9}$$

"Волновое число" ν, соответствующее "частоте" этой волны, определяется по формуле

$$\nu = \sqrt{s_1^2 + \cdots + s_n^2} \qquad \text{при } |s_k| \le \tfrac{m}{2} \text{ для } 1 \le k < n. \tag{10}$$

Согласно (5), никаких волн, кроме постоянной волны (с нулевой частотой), не должно появляться, если последовательность X_0, X_1, X_2, ... действительно случайная. Поэтому существование компонент с ненулевой частотой говорит об отклонении от случайности. Очень малые значения $f(s_1, \ldots, s_n)$ мало влияют на случайность. Например, если взять действительно случайную последовательность и изменить только каждый ее N-й член (при большом N), последовательность останется случайной, а значения $f(s_1, \ldots, s_n)$ будут порядка $1/N$; такими значениями можно пренебречь. Отметим, однако, что, согласно (8), значения $f(s_1, \ldots, s_n)$ равны либо 0, либо 1, а при $|f(s_1, \ldots, s_n)| = 1$ о случайности не может быть и речи. В свете этого интересно отметить, что *низкочастотные компоненты сильнее влияют на случайность, чем высокочастотные.* Посмотрим, например, что произойдет, если заменить X_k на $2X_k$ и m на $2m$. Согласно (4), $f(s_1, \ldots, s_n)$ не изменится, но теперь будут играть роль компоненты с $|s_k| \le m$, а не с $|s_k| \le m/2$. Изучение этой ситуации показывает, что если взять случайную последовательность целых чисел X_0, X_1, X_2, ... с $m = 2^e$ и начать отбрасывать младшие

разряды в двоичном представлении каждого члена последовательности, то при отбрасывании одного, двух, трех и т. д. двоичных разрядов будут появляться компоненты с волновыми числами 2^{e-1}, 2^{e-2}, 2^{e-3} и т. д. Этот факт, а также пример, приведенный в упр. (10), приводят к следующему интуитивному заключению. *Если ν_n—наименьшее ненулевое значение волнового числа (10), для которого $f(s_1, \ldots, s_n) \neq 0$ в линейной конгруэнтной последовательности с максимальным периодом, то последовательность X_0/m, X_1/m, X_2/m, ... можно считать последовательностью случайных чисел, равномерно распределенных между 0 и 1 и представленных с "точностью" или с "ошибкой округления" $1/\nu_n$, причем речь идет о независимости n последовательных значений с усреднением по полному периоду.* В упр. 27 содержится еще одно подтверждение этого принципа вместе с геометрической интерпретацией, показывающей более отчетливо важность значения ν_n.

Формула (8) дает "спектр" линейной конгруэнтной последовательности, т. е. показывает, какие типы волн имеются в преобразовании Фурье функции $F(t_1, \ldots, t_n)$. Мы видим, что $f(s_1, \ldots, s_n) = 0$ всегда, кроме случаев, когда

$$s_1 + s_2 a + s_3 a^2 + \cdots + s_n a^{n-1} \equiv 0 \pmod{m}, \tag{11}$$

и при этом $|f(s_1, \ldots, s_n)| = 1$. Следовательно, *для линейных конгруэнтных последовательностей с максимальным периодом наименьшее ненулевое волновое число в спектре равно*

$$\nu_n = \min \sqrt{s_1^2 + s_2^2 + \cdots + s_n^2}, \tag{12}$$

где минимум берется по всем n-наборам целых чисел $(s_1, s_2, \ldots, s_n) \neq (0, 0, \ldots, 0)$, удовлетворяющим (11). Отметим, что это условие не зависит от приращения c линейной конгруэнтной последовательности.

Одним из следствий всего сказанного является возможность определения верхнего предела случайности любой линейной конгруэнтной последовательности. Используя довольно глубокие теоретико-числовые методы, можно показать, что

$$\nu_n \leq \gamma_n m^{1/n}, \tag{13}$$

где γ_n принимает значения

$$1, (4/3)^{1/4}, 2^{1/6}, 2^{1/4}, 2^{3/10}, (64/3)^{1/12}, 2^{3/7}, 2^{1/2}$$

для $n = 1, 2, 3, 4, 5, 6, 7, 8$ [см. упр. 9, а также стр. 332 в книге J. W. S. Cassels, An Introduction to the Geometry of Numbers, Springer, Berlin (1959)]. Для любой периодической (с периодом m) последовательности $U_0, U_1, U_2, \ldots$ чисел, лежащих между 0 и 1, следовало бы ожидать границы, изменяющейся как $m^{1/n}$. Дело в том, что, согласно сформулированному выше правилу, выполнение с точностью $1/\nu$ предположения о независимости последовательных значений эквивалентно (при целом ν) тому, что каждое из ν^n возможных значений $(\lfloor \nu U_{k+1} \rfloor, \lfloor \nu U_{k+2} \rfloor, \ldots, \lfloor \nu U_{k+n} \rfloor)$ должно появляться приблизительно с одинаковой частотой в пределах периода; следовательно, из интуитивных соображений мы получаем приближенное неравенство $\nu^n \leq m$.

B. Примеры. Для большей ясности проиллюстрируем вышесказанное на примере. Возьмем линейную конгруэнтную последовательность с

$$X_0 = 0, \qquad a = 3141592621, \qquad c = 1, \qquad m = 10^{10}. \tag{14}$$

Минимальное ненулевое значение $s_1^2 + s_2^2$, для которого

$$s_1 + 3141592621 s_2 \equiv 0 \pmod{10^{10}},$$

соответствует $s_1 = 67654$, $s_2 = 226$, так что

$$\nu_2 = \sqrt{67654^2 + 226^2} \approx 67654.4.$$

Это означает, что последовательность

$$U_0, U_1, U_2, \ldots = X_0/m, X_1/m, X_2/m, \ldots$$

можно считать случайной в смысле независимости пар следующих друг за другом чисел (U_k, U_{k+1}) на полном периоде, если ограничиться точностью $1/67654$; т. е. первые 16 разрядов в двоичном представлении чисел можно считать случайными в указанном смысле. Аналогично, минимальное ненулевое значение $s_1^2 + s_2^2 + s_3^2$, для которого

$$s_1 + 3141592621 s_2 + 3141592621^2 s_3 \equiv 0 \pmod{10^{10}},$$

соответствует $s_1 = 227$, $s_2 = 983$, $s_3 = 130$; следовательно,

$$\nu_3 = \sqrt{1034718} \approx 1017.2.$$

Таким образом, если рассматривается независимость троек (U_k, U_{k+1}, U_{k+2}), мы имеем точность всего в 10 двоичных разрядов. Вероятно, 3141592621—не очень хороший множитель; позднее мы увидим, что он является приемлемым, но не лучшим (например, похожий множитель 3141592821 дает $\nu_3 \approx 1912$, но меньшее значение ν_2). Согласно (13), при $m = 10^{10}$ невозможно получить $\nu_3 > 2425$.

Минимальное ненулевое значение $s_1^2 + s_2^2 + s_3^2 + s_4^2$, для которого

$$s_1 + 3141592621 s_2 + 3141592621^2 s_3 + 3141592621^3 s_4 \equiv 0 \pmod{10^{10}},$$

имеет место при $s_1 = 52$, $s_2 = -203$, $s_3 = -54$, $s_4 = 125$, так что

$$\nu_4 = \sqrt{62454} \approx 249.9.$$

Требование независимости *четверок* понижает точность до 8 двоичных знаков (для большинства приложений этого вполне достаточно).

Значения ν_n для $n = 5, 6, \ldots$ менее важны, так как вряд ли независимость пятерок всегда действительно необходима. Например, при проверке серий (см. п. 3.3.2) редко учитываются даже четверки. (При рассмотрении средних по всему периоду, как в нашем случае, следует соблюдать осторожность, поэтому пренебрегать четверками в спектральном тесте не стоит; однако распределение пятерок вряд ли может понадобиться, во всяком случае при $m < 2^{40}$.) Для рассматриваемого датчика $s_1 = -8$, $s_2 = -14$, $s_3 = 6$, $s_4 = -18$, $s_5 = 34$ и

$$\nu_5 = \sqrt{1776} \approx 42.2,$$
$$\nu_6 = \sqrt{542} \approx 23.3.$$

Так как никто не знает, каковы наилучшие достижимые значения ν_n, трудно точно определить, какие значения ν_n можно считать удовлетворительными. Представляется разумным использовать в качестве меры случайности объем эллипсоида в n-мерном пространстве, определенного соотношением $(x_1 m - x_2 a - x_3 a^2 - \cdots - x_n a^{n-1})^2 + x_2^2 + \cdots + x_n^2 \le \nu_n^2$, так как этот объем пропорционален вероятности попадания в эллипсоид точек $(x_1, x_2, \ldots, x_n)$—*целочисленных* решений уравнений (11). Таким образом, для определения эффективности множителя a в линейной конгруэнтной последовательности с максимальным периодом мы предлагаем вычислять величину

$$C_n = \frac{\pi^{n/2} \nu_n^n}{(n/2)! m}. \tag{15}$$

(В этой формуле

$$\left(\frac{n}{2}\right)! = \left(\frac{n}{2}\right)\left(\frac{n}{2} - 1\right) \ldots \left(\frac{1}{2}\right)\sqrt{\pi} \qquad \text{для нечетных } n.) \tag{16}$$

Таким образом,

$$C_1 = 2\nu_1/m, \quad C_2 = \pi\nu_2^2/m, \quad C_3 = \frac{4}{3}\pi_3^3/m, C_4 = \frac{1}{2}\pi^2\nu_4^4/m \qquad \text{и т. д.}$$

Большие значения C_n соответствуют случайности, малые—отсутствию случайности. В табл. 1 представлены значения для некоторых типичных последовательностей (C_1 всегда равно 2). Датчики, представленные в строках 1–4 этой таблицы, были уже рассмотрены в п. 3.3.1 (см. рис. 2 и 5). У датчиков 1 и 2 слишком мал множитель. Очень плохой датчик 3 дает хорошее значение C_2, но очень плохие C_3 и C_4; для него $\nu_3 = 6$ и $\nu_4 = 2$. У датчика 4 "случайный" множитель; этот датчик успешно прошел испытания с использованием эмпирических тестов, но значения C_2, C_3 и C_4 у него не очень велики.

Датчик 7 мы только что рассматривали; рядом с ним представлены датчики с близкими параметрами. Отметим, что множитель 3141592221 приводит к аномально низкому значению C_3 (см. строку 5), однако при том же значении a с $m = 2^{35}$ (см. строку 9) получаются хорошие результаты.

Таблица 1

Некоторые результаты, полученные при помощи спектрального теста

Строка	a	m	C_2	C_3	C_4
1	23	10^8+1	0.000017	0.00051	0.014
2	2^7+1	2^{35}	0.000002	0.00026	0.040
3	$2^{18}+1$	2^{35}	3.14	0.000000002	0.000000003
4	3141592653	2^{35}	0.27	0.13	0.11
5	3141592221	10^{10}	1.35	0.06	4.67
6	3141592421	10^{10}	2.69	0.35	0.54
7	3141592621	10^{10}	1.44	0.43	1.91
8	3141592821	10^{10}	0.16	2.90	0.34
9	3141592221	2^{35}	1.24	1.69	1.11
10	3141592621	2^{35}	3.02	0.17	1.25
11	2718281821	2^{35}	2.59	1.15	1.75
12	$2^{23}+2^{12}+5$	2^{35}	0.015	2.78	0.066
13	$2^{23}+2^{13}+5$	2^{35}	0.015	1.48	0.066
14	$2^{23}+2^{14}+5$	2^{35}	1.12	1.66	0.066
15	$2^{22}+2^{13}+5$	2^{35}	0.75	0.30	0.066
16	$2^{24}+2^{13}+5$	2^{35}	0.0008	2.92	0.066
17	5^{13}	2^{35}	3.03	0.61	1.84
18	5^{15}	2^{35}	2.02	4.12	4.04
	Верхняя граница согласно (13):		3.63	5.90	9.86

У датчиков 12–16 в двоичном представлении a всего по 4 единицы; приемлемым можно считать только датчик 14, но и у него подозрительно низкое значение C_4. По странному совпадению все эти 5 датчиков дают одинаковое значение C_4; более того, во всех случаях $s_1 = -125$, $s_2 = 75$, $s_3 = 15$, $s_4 = 1$! Курьезным является также наличие у датчика 16 высокого значения C_3 при низком C_2; в этом случае $\nu_2 = \nu_3$, так как минимум при $n = 3$ достигается при $s_1 = -2043$, $s_2 = 2047$, $s_3 = 0$.

В датчиках 17 и 18 использованы множители, которые интенсивно применялись с тех пор, как их предложила О. Таусски в начале 50-х годов. По случайному совпадению наиболее популярный множитель 5^{15} дает наилучшие результаты из всех случаев, показанных в табл. 1.

Результаты, приведенные в табл. 1, и последующий опыт работы с многими из представленных здесь датчиков позволяют сказать, что множитель a *успешно прошел спектральный тест*, если каждое из C_2, C_3 и $C_4 \geq 0.1$; если все они больше (или равны) единицы, то можно считать, что спектральный тест пройден с блеском. Прежде чем рекомендовать датчик для общего пользования, можно вычислить также C_5, C_6 и т. д. Для того чтобы убедиться, что модуль m достаточно велик для обеспечения требуемой точности случайных чисел, надо анализировать значения ν_2, ν_3 и ν_4; при малых m удовлетворительные результаты, полученные с помощью спектрального теста, еще не гарантируют пригодности датчика для расчетов методом Монте-Карло с высоким разрешением.

C. Вывод вычислительного метода. Приведенные примеры иллюстрируют способы применения спектрального теста. Однако в наших рассуждениях остается, конечно, существенный пробел: существует ли хоть какая-нибудь возможность определить значение ν_n, затрачивая не слишком много машинного времени? Как, например, можно выяснить, что именно значениям $s_1 = 227$, $s_2 = 983$ и $s_3 = 130$ соответствует минимум суммы $s_1^2 + s_2^2 + s_3^2$ при соблюдении условия $s_1 + 3141592621 s_2 + 3141592621^2 s_3 \equiv 0 \pmod{10^{10}}$? Очевидно, что о простом переборе не может быть и речи.

Попытаемся отыскать подходящий вычислительный метод для решения этой задачи. Прежде всего перейдем от только что приведенной формулировки, опирающейся на формулы (11) и (12), к следующей эквивалентной задаче: определить минимум суммы

$$(x_1 m - a x_2 - a^2 x_3 - \cdots - a^{n-1} x_n)^2 + x_2^2 + x_3^2 + \cdots + x_n^2 \tag{17}$$

при целых значениях $x_1, x_2, \ldots, x_n$, из которых хотя бы одно не равно нулю.

Будет интересно и, вероятно, полезнее разработать численный метод решения более общей задачи: *определить минимум суммы*

$$(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n)^2 + \cdots + (a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n)^2 \tag{18}$$

при целых значениях $x_1, \ldots, x_n$*, из которых хотя бы одно не равно нулю,* и при условии, что матрица коэффициентов $A = (a_{ij})$ не вырождена. Выражение (18) называется "положительно определенной квадратичной формой".

В дальнейшем буквами x, y, ... будут обозначаться вектор-столбцы

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ldots$$

"Скалярное произведение" $x \cdot y = x_1 y_1 + \cdots + x_n y_n$ может быть записано в матричных обозначениях как $x^T y$, где T обозначает замену столбцов на строки, и наоборот (транспонирование). Для удобства введем следующие определения:

$$Q = A^T A, \quad B = A^{-1}, \quad R = Q^{-1} = BB^T. \tag{19}$$

Пусть A_j обозначает j-й *столбец* матрицы A, а B_i—i-ю *строку* матрицы B. Тогда имеем

$$B_i \cdot A_j = \delta_{ij}, \quad Q_{ij} = A_i \cdot A_j, \quad R_{ij} = B_i \cdot B_j. \tag{20}$$

Наша задача состоит в том, чтобы минимизировать (18), т. е. минимизировать $(Ax) \cdot (Ax) = x^T A^T Ax = x^T Qx$ для целых векторов $x \neq 0$.

Прежде всего сделаем задачу конечной, т. е. покажем, что не надо перебирать все бесконечное множество векторов x, чтобы найти минимум. Пусть e_k—вектор, k-я компонента которого равна 1, а остальные нулю. Тогда

$$x_k = e_k^T x = e_k^T BAx = (B^T e_k) \cdot (Ax) = B_k \cdot (Ax)$$

и, согласно неравенству Шварца,

$$(B_k \cdot (Ax))^2 \leq (B_k \cdot B_k)((Ax) \cdot (Ax)) = R_{kk}(x^T Qx).$$

Следовательно, если x—ненулевой вектор, минимизирующий $x^T Qx$, то

$$x_k^2 \leq R_{kk}(x^T Qx) \leq R_{kk}(e_j^T Qe_j) = R_{kk} Q_{jj}, \qquad 1 \leq j, k \leq n. \tag{21}$$

Это означает, что число векторов x, которые надо рассмотреть при поиске минимума, ограничено. На самом деле мы получили следующий более общий результат.

Лемма А. *Если*

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

—ненулевой целочисленный вектор с минимальным значением $x^T Qx$*, а* q*—значение* $y^T Q$ *для некоторого ненулевого целочисленного вектора* y*, то*

$$x_k^2 \leq R_{kk} q. \blacksquare \tag{22}$$

Очевидно, что правая часть (22) может быть все еще слишком большой, чтобы простой перебор был практически осуществим, так что требуется произвести дальнейшие усовершенствования. Обратимся к одному из наиболее простых и широко распространенных в математике приемов—методу замены переменных. Рассмотрим подстановку вида

$$y = Ux, \tag{23}$$

где U—целочисленная матрица, определитель которой $\det U = \pm 1$. Это означает, что если x—вектор-столбец, состоящий из целых чисел, то таков же и y, и обратно, если вектор y задан, то x можно определить из соотношения $x = U^{-1} y$. (Элементы матрицы U^{-1} будут целыми, так как она равна $\operatorname{adj}(U)/\det(U)$.) Следовательно, если в качестве x перебрать все целочисленные векторы, то это же множество значений пробежит и $y = Ux$, и обратно; далее, $y = 0$ только при $x = 0$. Поэтому можно перейти от задачи о минимизации $(Ax) \cdot (Ax)$ для целых $x \neq 0$ к эквивалентной задаче о нахождении минимума $(AU^{-1}y) \cdot (AU^{-1}y)$ для целых $y \neq 0$.

Лемма B. *Пусть U—любая целочисленная матрица с $\det U = \pm 1$, и пусть*

$$A' = AU^{-1}, \quad B' = Ub, \quad Q' = (U^{-1})^T Q U^{-1}, \quad R' = URU^T. \tag{24}$$

Задача минимизации, определенная матрицами A', B', Q', R', имеет то же самое решение, что и задача, определенная матрицами A, B, Q, R.∎

Теперь можно определить эффективный способ вычисления минимального значения следующим образом: перейти от исходной задачи к другой с помощью подходящей матрицы U и повторять это до тех пор, пока не получится задача, для которой неравенство в лемме A позволяет произвести полный перебор не слишком дорогой ценой.

Достаточно простой и пригодной для наших целей матрицей, определитель которой равен 1, может служить матрица

$$U = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ c_1 & \dots & c_{k-1} & 1 & c_{k+1} & \dots & c_n \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}$$

$$U^{-1} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ -c_1 & \dots & -c_{k-1} & 1 & -c_{k+1} & \dots & -c_n \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \tag{25}$$

где $c_1, \dots, c_n$—любые целые значения, k—фиксированное целое число. Все элементы матриц, кроме указанных, равны нулю. В этом случае соотношение $y = Ux$ означает просто, что $y_j = x_j$ для $j \neq k$ и $y_k = x_k + \sum_{j \neq k} c_j x_j$; безусловно, это наиболее естественный способ подстановки. Вычислить произведения, перечисленные в (24), в этом случае очень легко:

$$\begin{aligned} &A'_j = A_j - c_j A_k && \text{для } j \neq k,\ A'_k = A_k; \\ &B'_j = B_j && \text{для } j \neq k,\ B'_k = B_k + \textstyle\sum_{j \neq k} c_j B_j. \end{aligned} \tag{26}$$

Теперь нужно подобрать подходящие целые значения k и c_j. При любых c_j и целых k матрица U в (25) в принципе является вполне приемлемым преобразованием. В соответствии с (21) желательно выбрать целые числа $c_1, \dots, c_{k-1}, c_{k+1}, \dots, c_n$ так, чтобы *диагональные элементы* как матрицы Q', так и R' были как можно меньше. В связи с этим естественно возникают два следующих вопроса:

a) *Как лучше всего выбрать действительные числа c_j при $j \neq k$, чтобы минимизировать значения диагональных элементов матрицы $Q' = (U^{-1})^T Q U^{-1}$?*
b) *Как лучше всего выбрать действительные числа c_j при $j \neq k$, чтобы минимизировать значения диагональных элементов матрицы $R' = URU^T$?*

В случае (a) диагональные элементы матрицы Q', равные $A'_j \cdot A'_j$ согласно (20), будут изменены преобразованием U при всех $j \neq k$. Легко видеть, что минимум выражения

$$\begin{aligned} (A_j - c_j A_k) \cdot (A_j - c_j A_k) &= Q_{jj} - 2c_j Q_{jk} + c_j^2 Q_{kk} = \\ &= Q_{kk}(c_j - Q_{jk}/Q_{kk})^2 + Q_{jj} - Q_{jk}^2/Q_{kk} \end{aligned}$$

достигается при

$$c_j = Q_{jk}/Q_{kk}. \tag{27}$$

Геометрически (рис. 8) задача заключается в таком выборе коэффициента при векторе A_k, чтобы при вычитании $c_j A_k$ из вектора A_j результирующий вектор A'_j имел минимальную длину. Для этого надо выбрать такое c_j, чтобы A'_j было перпендикулярно к A_k (т. е. $A'_j \cdot A'_k = Q'_{jk} = 0$), а это достигается с помощью (27).

В случае (b) диагональные элементы матриц R' и R совпадают, кроме $R'_{kk} = B'_k \cdot B'_k$. Здесь нам надо выбрать c_j так, чтобы

Picture: Рис. 8. Геометрическая интерпретация вывода формулы (27).

вектор $B_k + \sum_{j \neq k} c_j B_j$ имел минимальную длину. Геометрически это означает, что мы добавляем к вектору B_k некоторый вектор, лежащий в $(n-1)$-мерной гиперплоскости, определяемой векторами $\{ B_j \mid j \neq k \}$. Так же, как в случае, показанном на рис. 8, минимум достигается, когда B'_k перпендикулярен гиперплоскости, B'_k перпендикулярен всем B'_j при $j \neq k$. Таким образом, необходимо решить систему уравнений $B'_k B'_j = 0$, т. е.

$$R_{kj} + \sum_{i \neq k} c_i R_{ij} = 0, \qquad 1 \leq j \leq n, j \neq k. \tag{28}$$

Строгое доказательство того, что задача (b) сводится к решению уравнений (28), рассмотрено в упр. 12.

Теперь, когда обе задачи (a) и (b) решены, мы пребываем в некотором недоумении: выбирать ли значения c в соответствии с (27), чтобы минимизировать диагональные элементы матрицы Q', или в соответствии с (28), чтобы минимизировать диагональные элементы R'? В обоих случаях правая часть (21) уточняется, поэтому неясно, какой вариант предпочтительнее. К счастью, ответ чрезвычайно прост: условия (27) и (28) совершенно одинаковы! Равенство $R' = (Q')^{-1}$ означает, что недиагональные элементы в k-й строке и k-м столбце Q' равны нулю тогда и только тогда, когда равны нулю недиагональные элементы в k-й строке и k-м столбце матрицы R'. Поэтому у задач (a) и (b) одно и то же решение. В результате оказывается, что можно сделать минимальными одновременно диагональные элементы Q и R. (Заметим, что мы только что открыли заново так называемый "процесс ортогонализации Шмидта".)

Конечно, на самом деле условия (a) и (b) выполняются при нецелых значениях c_j, а мы можем использовать в матрице U только целые значения. При этом сделать A'_j в точности перпендикулярным A'_k невозможно. Если, однако, выбрать в качестве c_j *ближайшие к* Q_{jk}/Q_{kk} *целые числа,* то это будет наилучшим целым решением задачи (a) и близким к (но *не* всегда равным) наилучшему решению задачи (b).

Производя преобразования вида (25) при разных k и при c_j, равных ближайшему целому к Q_{jk}/Q_{kk}, можно ожидать, что постепенно верхняя граница, определяемая выражением (21), спустится до уровня, при котором возможен полный перебор. На этом предположении основан приведенный ниже алгоритм. При написании настоящей главы автор провел несколько сотен машинных расчетов по этому алгоритму, причем сходимость оказывалась гораздо более быстрой, чем ожидалось. Достаточно было применить преобразование (25) всего 21 раз, чтобы в вариантах с $n = 6$ и с огромными значениями элементов матриц Q и R оставалось менее 500 случаев для прямого перебора. В результате расчеты на ЭВМ занимали всего несколько секунд.

D. Реализация спектрального теста. Приведем вычислительную процедуру, вытекающую из всего сказанного выше.

Алгоритм S. *(Спектральный тест.)* Спектральный тест применяется для оценки выбора множителя a в линейной конгруэнтной последовательности с максимальным периодом. (Вопрос о распространении этого теста на другие линейные конгруэнтные последовательности рассмотрен в упр. 20 и 21.) Кроме значения a задается также модуль m. Тест проверяет статистическую независимость последовательных отрезков из n чисел. Чаще всего тест применяется при $n = 2$, 3, 4 и иногда при несколько больших значениях n.

В алгоритме предполагается, что на входе заданы a, m и n; вычисляется $q = \nu_n^2$ (см. (12)). Используются $n \times n$-матрицы Q и R и вспомогательные n-мерные векторы X и c. Все операции с целыми числами должны производиться точно; для этого может потребоваться привлечение операций с повышенной точностью. Более подробно этот вопрос будет рассмотрен ниже.

S1 **[Начальная установка.]** Установить $X[1] \leftarrow 1$, $X[k+1] \leftarrow (aX[k]) \bmod m$ для $1 \le k < n$. Если какое-либо $X[k]$ больше $m/2$, установить $X[k] \leftarrow X[k] - m$. Затем сформировать матрицы

$$Q = \begin{pmatrix} m^2 & -mX_2 & -mX_3 & \dots & -mX_n \\ -mX_2 & 1+X_2^2 & X_2X_3 & \dots & X_2X_n \\ -mX_3 & X_2X_3 & 1+X_3^2 & \dots & X_3X_n \\ \vdots & & & & \vdots \\ -mX_n & X_2X_n & X_3X_n & \dots & 1+X_n^2 \end{pmatrix},$$
$$R = \begin{pmatrix} \sum X_j^2 & mX_2 & mX_3 & \dots & mX_n \\ mX_2 & m^2 & 0 & \dots & 0 \\ mX_3 & 0 & m^2 & \dots & 0 \\ \vdots & & & & \vdots \\ mX_n & 0 & 0 & \dots & m^2 \end{pmatrix}. \tag{29}$$

Для этого установить $Q[1,1] \leftarrow m^2$, $R[1,1] \leftarrow \sum_{1\le j\le n} X[j]^2$; для $1 < j \le n$ установить $Q[1,j] \leftarrow Q[j,1] \leftarrow -mX[j]$, $R[1,j] \leftarrow R[j,1] \leftarrow mX[j]$, $Q[j,j] \leftarrow 1 + X[j]^2$, $R[j,j] \leftarrow m^2$; и для $1 < j < k \le n$ установить $Q[j,k] \leftarrow Q[k,j] \leftarrow X[j]X[k]$, $R[j,k] \leftarrow R[k,j] \leftarrow 0$. **(Это те самые матрицы Q и R, которые фигурируют в (19), за исключением того, что $R = m^2Q^{-1}$, а не Q^{-1}, так что все последующие вычисления производятся с целыми числами.)**

Далее полагаем $k \leftarrow n$ и $q \leftarrow m^2$.

S2 **[Найти минимум Q_{jj}.]** Для $1 \le j \le n$, если $Q[j,j] < q$, установить $q \leftarrow Q[j,j]$.

S3 **[Переходить к перебору?]** Для $1 \le j \le n$ установить $c[j] \leftarrow \left\lfloor \sqrt{qR[j,j]}/m \right\rfloor$. Если теперь $\prod_{1\le j\le n}(2c[j]+1) \le 1000$, перейти к шагу **S6**. **(Согласно лемме A, для минимального решения $X[1], \dots, X[n]$ при любом j выполняется условие—$c[j] \le X[j] \le c[j]$. Смысл данного шага в том, чтобы перейти на S6, если для поиска точного минимума требуется перебрать не более 1000,—а на самом деле, как будет показано ниже,— менее 500 случаев. Конечно, нет необходимости вычислять точное значение $\prod(2c_j+1)$; при этом, кстати, возможно переполнение. Достаточно убедиться, превышает ли это значение 1000.)**

S4 **[Преобразовать.]** Для $1 \le j \le n$, $j \ne k$, присвоить $c[j]$ целые значения, как можно более близкие к $Q[j,k]/Q[k,k]$, например $\lfloor Q[j,k]/Q[k,k] + \frac{1}{2} \rfloor$ **(см. (27))**. Затем для $1 \le j \le n$, $j \ne k$ и $c[j] \ne 0$ выполнить операции $\texttt{TRANS}(Q, j, k, -c[j])$ и $\texttt{TRANS}(R, k, j, c[j])$.

Операция $\texttt{TRANS}(P, i, j, t)$ над матрицей P определяется следующим образом. Для $1 \le r \le n$ установить $P[i,r] \leftarrow P[i,r] + tP[j,r]$; *затем* для $1 \le r \le n$ установить $P[r,i] \leftarrow R[r,i] + tP[r,j]$. **[Таким образом, строка j умножается на t и прибавляется к строке i, затем столбец j умножается на t и прибавляется к столбцу i. В результате на шаге S4 производится преобразование (24) с использованием матрицы U вида (25).]**

S5 **[Изменить k.]** Уменьшить k на 1; затем, если $k = 0$, установить $k \leftarrow n$. Перейти на **S2**.

S6 **[Подготовка к перебору.]** **(Теперь будет определяться абсолютный минимум с помощью прямого перебора.)** Установить $k \leftarrow n$ и $X[j] \leftarrow 0$ для $1 \le j \le n$.

S7 **[Перейти к следующему значению $X[k]$.]** Установить $X[k] \leftarrow X[k] + 1$. Если $X[k] > c[k]$, перейти на **S9**.

S8 **[Перейти к следующему k.]** Установить $k \leftarrow k + 1$. Если $k \le n$, установить $X[k] \leftarrow -c[k]$ и повторить шаг **S8**. Если $k > n$, установить $q' \leftarrow \sum_{1\le i\le n}\sum_{1\le j\le n} X[i]X[j]Q[i,j]$, и, если $q' < q$, установить $q \leftarrow q'$.

S9 **[Уменьшить k.]** Установить $k \leftarrow k - 1$. Если $k \ge 1$, вернуться на **S7**, в противном случае выполнение алгоритма заканчивается. **(Замечание: шаги S6–S9 являются простейшим случаем метода обратного перебора, который подробно описан в гл. 7.)** ∎

Определенная с помощью этого алгоритма величина q позволяет вычислить $\nu_n = \sqrt{q}$ и определить C_n по формуле **(15)**. Значение величин ν_n и C_n и условия, при выполнении которых можно считать, что множитель a прошел испытание удовлетворительно, обсуждались выше в подпункте B.

Так как обычно m—это размер слова машины, на которой производится реализация алгоритма S, необходимо при расчетах по этому алгоритму пользоваться арифметикой над целыми числами с увеличенной точностью. Опыт показывает, что достаточно тройной точности (когда каждое целое число размещается в трех ячейках). Например, при подготовке табл. 1 автор использовал 76-разрядную арифметику. Этого было достаточно при $m = 10^{10}$ и $n \ge 6$, однако при $m = 2^{35}$ во многих случаях происходило переполнение. По-видимому, почти для всех вариантов с $m = 2^{35}$ будет достаточно примерно 90 двоичных разрядов. Каких-либо теоретических данных относительно величины промежуточных результатов пока не существует.

На практике алгоритм S оказывается удивительно эффективным; при $m = 2^{35}$ число повторений шагов S2–S5 оказывалось равным 6, 10, 14, 17, 21 для $n = 2, 3, 4, 5, 6$. Однако до сих пор нет теории, полностью описывающей алгоритм, и на самом деле в некоторых случаях *метод может просто зациклиться.* Так что термин "алгоритм" в данном случае, строго говоря, неприменим. Легко изменить эту процедуру так, чтобы число итераций было ограничено (см. упр. 16). Достаточно ввести еще одну переменную d и устанавливать $d \leftarrow 0$ на шаге S1 и в операции TRANS. Кроме того, в конце шага S3 добавить: "$d \leftarrow d+1$; если $d > n$, перейти на S6". Однако, возможно, что это приведет к очень длинному перебору на шагах S6–S9, как показано в упр. 18. В такой ситуации значительными преимуществами обладают приемы, которые обсуждаются в упр. 22 и 23. Алгоритм зацикливается при $n = 2$, $a = 1025$, $m = 2^{46}$, хотя подобная неудача бывает редко.

В одном из просчитанных автором случаев при $n = 6$ "ожидаемая длина перебора" $\prod(2c_j + 1)$ на шаге S3 принимала последовательно следующие значения:

$$1 \times 10^{43}, 6 \times 10^{42}, 2 \times 10^{42}, 9 \times 10^{41}, 2 \times 10^{41}, 6 \times 10^{33}, 4 \times 10^{33},$$
$$1 \times 10^{29}, 1 \times 10^{20}, 6 \times 10^{19}, 4 \times 10^{18}, 9 \times 10^{12}, 4 \times 10^{10}, 3 \times 10^{8},$$
$$1 \times 10^{8}, 8 \times 10^{7}, 1 \times 10^{7}, 7 \times 10^{6}, 1.7 \times 10^{7}, 1.8 \times 10^{7}, 7 \times 10^{5},$$
$$1 \times 10^{5}, 5 \times 10^{4}, 3825, 3825, 675.$$

Таким образом, эта величина уменьшается от 10^{43} до значения ниже 1000, причем не монотонно: дважды значение *увеличивается.* Вероятно, дальнейшие итерации еще больше снизят значение 675, так что, по-видимому, константу 1000 в шаге S3 следует уменьшить, скажем, до 100. (*Замечание.* Истинное число наборов $X[1], \ldots, X[n]$, испытанных в полном переборе (шаги S6–S9), равно лишь $\lfloor \frac{1}{2} \prod(2c_j + 1) \rfloor$, а не $\prod(2c_j + 1)$, так как алгоритм имеет дело только с такими векторами, у которых первый ненулевой элемент положителен.)

Рассмотрим кратко пример алгоритма S в действии, когда $a = 3141592621$, $m = 10^{10}$, $n = 3$. В табл. 2 в первых строках представлены Q и R, приготовленные на шаге S1. Исследование этих матриц средствами леммы A потребует проверки 10^{29} случаев, что выходит за всякие границы. После шести итераций на шагах S2–S5 элементы матриц Q и R стали намного меньше (см. строку 7 табл. 2), и, согласно лемме A, теперь в этой новой задаче $|x_1| \leq 3$, $|x_2| \leq 3$, $|x_3| \leq 14$. Дальнейшее уменьшение с помощью леммы B приводит нас к строке 8: в матрице Q (строка 7) прибавляем столбец 3 к столбцу 2, строку 3 к строке 2, затем 3 раза вычитаем столбец 3 из столбца 1 и также три раза строку 3 из строки 1. В матрице R вычитаем столбец 2 из столбца 3, затем вычитаем строку 2 из строки 3, потом прибавляем три раза столбец 1 к столбцу 3, а строку 1 снова три

Таблица 2

Пример алгоритма S

Строка	Матрица Q		
1.	1 00000 00000 00000 00000	− 31415 92621 00000 00000	36783 50359 00000 00000
	− 31415 92621 00000 00000	9869 60419 63216 49642	− 11555 87834 52871 00939
	36783 50359 00000 00000	− 11555 87834 52871 00939	13530 26136 35554 28882
⋮		⋮	
7.	1160 62418	− 110 45623	324 06810
	− 110 45623	189 42062	− 70 72864
	324 06810	− 70 72864	99 86024
8.	114 95774	126 21707	24 48738
	126 21707	147 82358	29 13160
	24 48738	29 13160	99 86024

Строка	Матрица R		
1.	23399 86555 98770 78523	31415 92621 00000 00000	− 36783 50359 00000 00000
	31415 92621 00000 00000	1 00000 00000 00000 00000	0
	− 36783 50359 00000 00000	0	1 00000 00000 00000 00000
⋮		⋮	
7.	13913 04805 78992	− 11890 71034 30888	− 53572 76149 67948
	− 11890 71034 30888	10880 07572 69932	46294 02921 32522
	− 53572 76149 67948	46294 02921 32522	2 07645 57301 67787
8.	13913 04805 78992	− 11890 71034 30888	57 09301 99916
	− 11890 71034 30888	10880 07572 69932	− 258 17754 30074
	57 09301 99916	− 258 17754 30074	1062 71591 61243

раза прибавляем к строке **3**. Это уменьшает Q и R, так что, согласно лемме **A**, теперь осталось проверить значение $|x_1| \le 3$, $|x_2| \le 3$, $|x_3| \le 1$, чтобы найти абсолютный минимум. В действие приводится метод перебора на шагах **S6**–**S9**, который находит комбинацию $x_1 = 1$, $x_2 = -1$, $x_3 = 0$, реализующую минимальное значение $x^TQx = 1034718$. Эти вычисления можно было бы сделать с помощью настольной вычислительной машинки за несколько часов, хотя в начале задача выглядела весьма внушительно.

Спектральный тест впервые появился в статье P. Ковэю и P. Макферсона (R. R. Coveyou, R. D. MacPherson, Fourier Analysis of Uniform Random Number Generators, *JACM,* **14** (1967), 100–119). В этой статье описан алгоритм, в сущности подобный алгоритму S, за исключением несколько отличного правила преобразования на шаге **S4**.

Упражнения

1. **[M20]** Выведите соотношение **(2)** из **(1)**.
2. **[M20]** Предполагая, что $0 \le s_1 \ldots, s_n < m$, выведите соотношение **(1)** из **(2)**.
3. **[M22]** (a) Пусть $F(t)$ определена для целых значений t, $0 \le t < m$, и область ее определения расширена на все целые числа с помощью формулы $F(t) = F(t \bmod m)$. Пусть $f(s)$—преобразование Фурье функции $F(t)$, определенное в соответствии с **(1)** для $n = 1$. Найдите преобразование Фурье функции $F(t+1)$, выразив его через $f(s)$.

   (b) Найдите преобразование Фурье суммы $\sum_{0\le k<m} F(k)G(t-k)$, выразив его через преобразования Фурье F и G.

>4. **[M22]** Пусть X_0, X_1, X_2, ...—периодическая последовательность целых чисел с периодом m, такая, что $0 \le X_k < m$, и пусть $f(s_1, s_2)$—функция, определенная формулой **(4)**. Выразите через f вероятность того, что $X_{k+1} < X_k$ в этой последовательности. (Упростите насколько возможно ваш ответ, чтобы явно были представлены коэффициенты при каждом частном значении f.)
5. **[M23]** Пусть $X_0, X_1, X_2, \ldots$—периодическая последовательность целых чисел с длиной периода m и такая, что $0 \le X_k < m$. Выразите через функцию F, определяемую соотношением **(3)**, и функцию f, определяемую соотношением **(4)**, сумму $\sum_{0\le k<m} X_kX_{k+1}$, представляющую важную часть формулы для коэффициента последовательной корреляции (формула **(3.3.2-23)**).

>6. **[M28]** Докажите теорему **3.3.3P**, используя **(8)** и результат упр. **4**.
7. **[M40]** Выведите формулы, позволяющие достаточно эффективно вычислять точное значение вероятности того, что $X_{k+2} < X_{k+1} < X_k$ для линейной конгруэнтной последовательности с максимальным периодом. [*Указание.* С помощью методов, используемых в упр. **6**, и результата упр. **3.3.3-20** эта величина может быть явно выражена через обобщенные суммы Дедекинда. Заметьте, что задача сильно усложняется, если не пользоваться преобразованием Фурье.]
8. **[M22]** Найдите приемлемо простое выражение суммы $\sum_{0\le k<m} X_kX_{k+1}$ для линейной конгруэнтной последовательности с максимальным периодом, используя результат упр. **5**, а также формулу **(8)**.
9. **[BM30]** (Ш. Эрмит.) Задана $n \times n$-матрица A. Докажите, что существует ненулевой целочисленный вектор x, такой, что $Ax \cdot Ax \le \left(\frac{4}{3}\right)^{(n-1)/2} (\det A)^{2/n}$. [*Указание.* Сначала показать, что для всех $\varepsilon > 0$ существует целочисленная матрица U, детерминант которой равен 1, и произведение $AUx \cdot AUx$ находится в ε-окрестности максимальной нижней границы ее значений при $(x_1, x_2, \ldots, x_n) = (1, 0, 0, \ldots, 0)$. Затем доказать общее утверждение индукцией по n, записав $Ax \cdot Ax$ в виде $\alpha(x_1 + \beta_2x_2 + \cdots + \beta_nx_n)^2 + g(x_2, \ldots, x_n)$, где g соответствует $(n-1) \times (n-1)$-матрице A'.]
10. **[BM30]** (Ковэю и Макферсон). Пусть X_0, X_1, X_2, ...—последовательность целых чисел, лежащих в пределах $0 \le X_k < m$; ей соответствует преобразование Фурье $f(s_1, \ldots, s_n)$, определенное формулой **(4)**. Пусть $U_k = X_k/m$, а V_0, V_1, V_2, ...—последовательность независимых истинно случайных действительных чисел, равномерно распределенных между 0 и 1; λ—число, меньшее 1, и $W_k = (U_k + \lambda V_k) \bmod 1$. (Теперь W_k—это случайным образом "размазанные" в пределах λ члены последовательности U_k.) Определим коэффициенты Фурье любой последовательности действительных чисел между 0 и 1 как

$$\lim_{N\to\infty} \frac{1}{N} \sum_{0\le k<N} \exp(2\pi i(s_1U_k + \cdots + s_nU_{k+n-1})).$$

    Выразите коэффициенты Фурье последовательности $\langle W_k \rangle$ через коэффициенты Фурье последовательности $\langle X_k \rangle$.
11. **[M10]** Уравнение **(8)** показывает, что значение c в линейной конгруэнтной последовательности с максимальным периодом не влияет на коэффициенты Фурье, а изменяет лишь "аргумент" комплексного числа $f(s_1, \ldots, s_n)$. Другими словами, абсолютное значение $f(s_1, \ldots, s_n)$ не зависит от c.

Но можно ли выбрать c так, чтобы неслучайный эффект одной волны $f(s_1, \ldots, s_n)$ уничтожался бы "противоположным" эффектом другой волны $f(s'_1, \ldots, s'_n)$?

12. **[BM23]** Докажите, не используя геометрических аргументов, что любое решение "проблемы (b)", сформулированной в подпункте C текста, должно быть одновременно решением системы уравнений (28).
13. **[BM30]** В тексте остался в тени довольно важный вопрос: было сделано молчаливое предположение, что, если A—произвольная невырожденная матрица действительных чисел, функция (18) *имеет* минимум, который *достигается* на некотором целочисленном векторе x.

(a) Докажите, что наибольшая нижняя граница величины (18), взятая по всем ненулевым целочисленным векторам x, достигается при некотором x, если A—невырожденная матрица.

(b) Покажите, что, если A—вырожденная матрица, такого ненулевого целочисленного вектора, на котором достигается наибольшая нижняя граница (18), может не существовать.

>14. **[24]** Выполните вручную алгоритм S для $m = 100$, $a = 41$, $n = 3$. Замените константу "1000" на шаге S3 числом "3".
15. **[M18]** Что произойдет, если операцию "$k \leftarrow n$" в конце шага S1 заменить на "$\leftarrow 1$"?
16. **[M25]** Не исключено (хотя этого еще не наблюдалось), что алгоритм S может зациклиться, повторяя бесконечное число раз шаги S2–S5. Покажите, что это может произойти тогда и только тогда, когда при последовательном выполнении n раз шага S4 не происходит преобразований (т. е. нет операций TRANS).

>17. **[M28]** Модифицируйте алгоритм S так, чтобы кроме вычисления q в нем определялся бы набор целых чисел $s_1, \ldots, s_n$, удовлетворяющих (11) при $s_1^2 + \cdots + s_n^2 = q$. [*Указание.* Алгоритм S сохраняет только значения Q и R из (19), но не A и B. Если сохранить значения A и/или B при выполнении алгоритма, по-видимому, не слишком трудно будет получить значения $s_1, \ldots, s_n$.]
18. **[M25]** Найдите 3×3-матрицу A, такую, что если $Q = A^TA$ и $R = Q^{-1}$, то шаги S2–S5 алгоритма S никогда не закончатся переходом к шагу S6 (так что вычисления никогда не прекратятся). [*Указание.* Рассмотреть "комбинаторные матрицы", т. е. матрицы, элементы которых имеют вид $a + b\delta_{ij}$; ср. с упр. 1.2.3-39.]
19. **[M20]** Покажите, что последовательность Фибоначчи $\bmod m$ служит плохим источником случайных чисел, убедившись, что соответствующая функция $f(s_1, s_2, s_3)$, определенная в (4), имеет большие низкочастотные компоненты.

>20. **[M24]** Вычислите коэффициенты Фурье $f(s_1, \ldots, s_n)$ для линейной конгруэнтной последовательности, определенной величинами $X_0 = 1$, $c = 0$, $m = 2^e \geq 8$ и $a \bmod 8 = 5$. Обсудите, как обобщить спектральный тест для этого типа датчиков случайных чисел (определенный в тексте только для линейных конгруэнтных последовательностей с *максимальным* периодом, тогда как у определенной здесь последовательности длина периода равна $m/4$).
21. **[M25]** Проделайте предыдущее упражнение, считая, что $a \bmod 8 = 3$.

>22. **[M30]** Постройте алгоритм, подобный алгоритму S, за исключением того, что в нем используется преобразование U, для которого все ненулевые недиагональные элементы находятся в *столбце* k, а не в *строке* k, как в (25). Сравните этот метод с алгоритмом S. Покажите, что в этом алгоритме величина $\prod(2c_j + 1)$ в шаге S3 никогда не увеличивается от одной итерации к другой.
23. **[M50]** Несмотря на то что в примере из упр. 18 основной цикл алгоритма S вынужден бесконечно повторяться, для "двойственного" алгоритма из упр. 22 тот же пример не представляет никакой проблемы. Для удобства будем называть последний метод алгоритмом S'. Так как алгоритм S' в сущности является алгоритмом S, в котором поменялись ролями матрицы Q и R, также существуют матрицы, заставляющие зацикливаться и этот алгоритм. Отсюда вытекает идея комбинации двух методов. Например, мы можем использовать алгоритм S, пока он не застрянет, затем переключиться на алгоритм S' до тех пор, пока *этот* не застрянет, вернуться затем снова к алгоритму S и т. д.

Пользуясь таким комбинированным алгоритмом и игнорируя ветвление на шаге S3, мы можем при реализации вычислительной процедуры оказаться в одной из двух ситуаций: (a) устанавливается цикл, в котором, каждый из алгоритмов S и S' некоторым образом попеременно преобразуют Q и R, так что вычисления никогда не смогут прекратиться, или (b) мы получаем матрицы Q и R, на которые не влияют ни алгоритм S, ни алгоритм S'.

Эти наблюдения, естественно, приводят к постановке следующих трех вопросов, на которые должен быть дан ответ, если нам нужно иметь полностью удовлетворительное решение вычислительной проблемы, сформулированной в этом разделе. Может ли реализоваться случай (a)? Как долго можно искать константы $\prod(2c_j + 1)$ на шаге S3 в случае (b)? Существует ли общая процедура вычисления $\min\{\, x^TQx \mid \text{целые } x \neq 0 \,\}$ для положительно определенной матрицы Q, которая была бы лучше, чем только что описанная комбинация алгоритмов S и S'?

Замечание. Второй из поставленных выше вопросов можно свести к следующему: *Пусть* Q*— симметричная положительно определенная* $n \times n$*-матрица действительных чисел, у которой диагональные элементы равны* 1, *а* $|q_{ij}| \le 1/2$ *для* $i \ne j$. *Пусть* $R = Q^{-1}$, *а* $|r_{ij}| \le (1/2)r_{jj}$ *для* $i \ne j$. *Как велико может быть при этих условиях число* r_{11}*?* До сих пор не встречались подобные матрицы, для которых $r_{11} \ge 2$. Если принять все недиагональные элементы матрицы Q равными $-1/n$, то можно найти, что $r_{11} = 2n/(n+1)$. Этот пример показывает, что никогда нельзя предполагать, что константа на шаге S3 примет значение, меньшее 3^n, для произвольной положительно определенной матрицы, даже в случае комбинации алгоритмов S и S'. В примере достигается $\max r_{11}$ для $n = 2$, но не для $n = 3$.

24. [M20] Сравните преобразование Фурье $f(s_1, \ldots, s_n)$, заданное формулой (1), с *производящей функцией* от n переменных для $F(t_1, \ldots, t_n)$, определенной обычным образом:

$$g(z_1, \ldots, z_n) = \sum_{0 \le t_1 \ldots t_n < m} F(t_1, \ldots, t_n) z_1^{t_1} \ldots z_n^{t_n}.$$

25. [BM28] (Р. Ковэю.) Проанализируйте двумерное преобразование Фурье $f(s_1, s_2)$ для последовательности (3.2.2-4) в модифицированном методе середины квадрата. [*Указание:* рассмотреть двойную сумму для $|f(s_1, s_2)|^2 = f(s_1, s_2) \times f(-s_1, -s_2)$.]

>26. [M26] Чему равно значение трехмерного преобразования Фурье $f(s_1, s_2, s_3)$ последовательности (3.2.2-8), имеющей длину периода $p^2 - 1$, в случае $k = 2$?

>27. [BM24] (Дж. Марсалья.) Пусть X_0, a, m, c порождают линейную конгруэнтную последовательность, и U_0, U_1, U_2, $\ldots = X_0/m$, X_1/m, X_2/m $\ldots$. Предположив, что s_1, s_2, $\ldots$, s_n—целые" числа, удовлетворяющие (11), докажите следующие утверждения: (a) каждый набор n чисел $(U_k, U_{k+1}, \ldots, U_{k+n-1})$ определяет в n-мерном пространстве точку, лежащую на одной из гиперплоскостей, задаваемых уравнением

$$s_1 x_1 + s_2 x_2 + \cdots + s_n x_n = N + (s(a) - s(1))c/(a-1)m,$$

где N—целое, а $s(a)$ определяется формулой (7). (b) Расстояние между соседними плоскостями равно $1/\sqrt{s_1^2 + \cdots + s_n^2}$. (c) Число таких гиперплоскостей, пересекающих n-мерный куб $0 \le x_1, \ldots, x_n < 1$, равно $|s_1| + \cdots + |s_n| - \delta$, где $\delta = 1$, если $s_i s_j < 0$ для некоторых i, j, в противном случае $\delta = 0$.

Замечание. По словам Гивенса, последовательные случайные векторы, вырабатываемые линейными конгруэнтными методами, "остаются главным образом в плоскостях". Это упражнение убедительно демонстрирует неудобство линейных конгруэнтных последовательностей для применений в методе Монте-Карло, если требуется высокое разрешение. Например, трехмерные векторы (U_k, U_{k+1}, U_{k+2}), вырабатываемые датчиком (14), все лежат в параллельных плоскостях, отстоящих друг от друга на $1/\nu_3 \approx 0.001$ единиц.

3.4. ДРУГИЕ ВИДЫ СЛУЧАЙНЫХ ВЕЛИЧИН

Мы уже знаем, как с помощью вычислительной машины вырабатывать последовательность чисел U_0, U_1, U_2, $\ldots$, которые ведут себя так, как если бы их выбирали случайно и независимо из множества чисел, равномерно распределенных между нулем и единицей. Для практических применений часто бывают нужны случайные величины с другими распределениями. Например, если мы хотим сделать выбор между k возможными альтернативами, нам могут понадобиться случайные *целые числа* от 1 до k. Если для некоторого моделирования требуются случайные промежутки времени между какими-то независимыми событиями, пользуются случайными величинами с "экспоненциальным распределением". Иногда мы вообще не нуждаемся в случайных *числах,* а интересуемся случайными перестановками (расстановкой n элементов в случайном порядке) или случайными сочетаниями (т. е. случайным выбором k элементов из n возможных).

В принципе любую из этих случайных величин можно получить с помощью случайных равномерно распределенных чисел $U_0, U_1, U_2, \ldots$ [9]. Для этого существует ряд важных приемов, эффективно используемых при работе на вычислительных машинах, Изучение этих приемов позволяет понять, как можно применять случайные числа в любых приложениях метода Монте-Карло.

Можно допустить, что когда-нибудь кто-нибудь изобретет датчик, вырабатывающий эти другие случайные величины *непосредственно,* а не с помощью случайных чисел с равномерным распределением. Кроме случая датчика "случайных битов", описанного в п. 3.2.2, до сих пор не было доказано, что какой-либо подобный прямой метод может оказаться более выгодным.

---

[9] Ниже везде, где говорится просто о случайных числах, подразумеваются случайные числа с равномерным распределением в интервале $(0, 1)$.—*Прим. перев.*

В следующем разделе будем предполагать, что U_0, U_1, U_2, ...—случайная последовательность действительных чисел, равномерно распределенных между нулем и единицей. Буквой U без индекса будем обозначать текущий элемент этой последовательности. Будем считать, что новое U вырабатывается независимо от того, понадобится ли оно нам или нет. Обычно случайные числа представляются содержимым всего слова вычислительной машины, в предположении, что десятичная точка расположена слева. Конечно, на самом деле число в машине представляется с ограниченной точностью. И если эта точность по какой-либо причине окажется недостаточной, всегда можно объединить несколько U водно число с более высокой точностью.

3.4.1. Числовые распределения

В этом разделе рассказывается о наиболее известных методах получения случайных величин для различных важных распределений. Многие из этих методов первоначально были предложены Джоном фон Нейманом в начале 50-х годов, а затем постепенно улучшались другими людьми, особенно Джорджем Марсальей.

A. Случайный выбор из конечного множества. Простейшим и наиболее общим типом распределений, необходимых для практики, являются распределения *целочисленных* случайных величин. Целое число от 0 до 7 можно извлечь из 3 битов слова U в двоичной вычислительной машине. В этом случае его следует брать из *самых старших* (левых) разрядов слова, так как у многих датчиков младшие разряды недостаточно случайны. (Это обсуждалось в п. 3.2.1.1.)

Вообще, чтобы получить случайное число X между 0 и $k-1$, мы можем *умножить* U на k и положить $X = \lfloor kU \rfloor$. Для MIX можно написать

```
01 LDA  U
02 MUL  K
```
(1)

После того как эти две команды будут выполнены, искомое целое число окажется в регистре A. Если требуется случайное целое между 1 и k, мы добавляем к результату единицу. [За (1) последует команда "INCA 1".]

Этот метод дает каждое целое число с равной вероятностью. [Небольшой ошибкой, вызванной конечностью размера машинного слова (см. упр. 2), вполне можно пренебречь, если k мало, например если $k/m < 1/10000$.] В более общем случае мы могли бы захотеть приписать различным целым числам разные веса. Предположим, что значение $X = x_1$ должно получаться с вероятностью p_1, $X = x_2$ с вероятностью p_2, ..., и $X = x_k$ с вероятностью p_k. Мы можем выработать равномерно распределенное число U и принять

$$X = \begin{cases} x_1, & \text{если } 0 \le U < p_1; \\ x_2, & \text{если } p_1 \le U < p_1 + p_2; \\ \dots & \\ x_k, & \text{если } p_1 + p_2 + \cdots + p_{k-1} \le U < 1. \end{cases} \tag{2}$$

(Заметим, что $p_1 + p_2 + \cdots + p_k = 1$.)

Существует оптимальный способ сравнения U с различными значениями $p_1 + p_2 + \cdots + p_s$, как это требуется для вычислений (2); это обсуждается в п. 2.3.4.5. Можно воспользоваться также командой "поиск в таблице", имеющейся в некоторых машинах. К частным случаям применимы более эффективные методы. Например, чтобы получить одно из одиннадцати значений 2, 3, ..., 12 с соответствующими вероятностями $\frac{1}{36}$, $\frac{2}{36}$, ..., $\frac{6}{36}$, ..., $\frac{2}{36}$, $\frac{1}{36}$, могли бы вычислить два независимых случайных целых числа от 1 до 6, а затем их сложить.

Однако ни один из перечисленных выше способов не позволяет максимально быстро выбрать $x_1, \dots, x_k$ с правильными вероятностями. Значительно более эффективный для большинства случаев метод, требующий небольшого увеличения памяти, рассматривается в упр. 20 и 21.

B. Общие методы для непрерывных распределений. Самое общее распределение действительных случайных величин описывается в терминах "функции распределения" $F(x)$; мы требуем, чтобы случайная величина X принимала значение, меньшее или равное x, с вероятностью $F(x)$:

$$F(x) = \text{вероятность}(X \le x). \tag{3}$$

Эта функция всегда монотонно увеличивается от нуля до единицы:

$$F(x_1) \le F(x_2), \qquad \text{если } x_1 \le x_2;\ F(-\infty) = 0,\ F(+\infty) = 1. \tag{4}$$

Примеры функций распределения приводятся в п. 3.3.1, рис. 3. Если $F(x)$ непрерывна и строго возрастает (так что $F(x_1) < F(x_2)$, если $x_1 < x_2$), она принимает все значения между нулем и единицей, и существует обратная функция $F^{-1}(y)$, такая, что если $0 < y < 1$, то

$$y = F(x) \text{ тогда и только тогда, когда } x = F^{-1}(y). \tag{5}$$

Общий способ вычисления случайной величины X с непрерывной строго возрастающей функцией распределения $F(x)$ заключается в том, что полагают

$$X = F^{-1}(U). \tag{6}$$

В самом деле, вероятность того, что $X \le x$, является вероятностью того, что $F^{-1}(U) \le x$, т. е. вероятностью события $U \le F(x)$, а она равна $F(x)$.

Теперь задача сводится к одной из проблем численного анализа для определения хороших методов вычисления $F^{-1}(U)$ с требующейся точностью. Численный анализ выходит за рамки этой книги. Однако существует ряд важных рациональных приемов, пригодных для ускорения этого общего метода, и мы их здесь рассмотрим.

Прежде всего, если X_1 и X_2—независимые случайные величины с функциями распределения $F_1(x)$ и $F_2(x)$, то

$$\begin{aligned} &\max(X_1, X_2) \qquad \text{имеет распределение } F_1(x)F_2(x), \\ &\min(X_1, X_2) \qquad \text{имеет распределение } F_1(x) + F_2(x) - F_1(x)F_2(x). \end{aligned} \tag{7}$$

(См. упр. 4.) Например, случайное число U имеет функцию распределения $F(x) = x$ для $0 \le x < 1$. Если $U_1, U_2, \ldots, U_t$—независимые случайные числа, то $\max(U_1, U_2, \ldots, U_t)$ имеет функцию распределения $F(x) = x^t$, $0 \le x < 1$. Это лежит в основе теста "наибольшее из t", описанного в п. 3.3.2. Заметим, что обратная функция в этом случае есть $F^{-1}(y) = \sqrt[t]{y}$. Таким образом, в частном случае $t = 2$ мы видим, что две формулы

$$X = \sqrt{U} \text{ и } X = \max(U_1, U_2) \tag{8}$$

приводят к эквивалентным распределениям случайной величины X, хотя на первый взгляд это не очевидно. Мы избавлены от необходимости вычислять квадратный корень из случайной величины.

Таких приемов бесконечно много. Любой алгоритм, которому на входе задаются случайные числа, на выходе выдает случайную величину с *некоторым* распределением. Задача заключается в определении общих методов построения алгоритма, имея заданную функцию распределения на выходе.

Пользуясь соотношением (7), можно получить *произведение* двух функций распределения. Существует также метод *смешивания* двух распределений: предположим, что

$$F(x) = pF_1(x) + (1-p)F_2(x), \qquad 0 < p < 1. \tag{9}$$

Мы можем вычислить значение случайной величины X с распределением $F(x)$, определив сначала случайное число U. Если $U < p$, считаем, что X имеет распределение $F_1(x)$, если же $U \ge p$, то X—случайная величина с распределением $F_2(x)$.

Эта процедура может быть очень полезна, если p близко к единице, а $F_1(x)$—распределение, которое легко моделировать. Тогда, несмотря на то что выработка случайных значений по распределению $F_2(x)$ может быть более трудоемкой, чем для требующегося полного распределения $F(x)$, более трудные вычисления должны будут проводиться редко, с вероятностью $(1-p)$. Ниже мы увидим, что эта идея с успехом используется для нескольких важных случаев.

C. Нормальное распределение. *Нормальное распределение со средним значением, равным нулю, и стандартным отклонением, равным единице,* является, возможно, важнейшим из неравномерных непрерывных распределений:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt. \tag{10}$$

Для выработки нормально распределенных случайных величин существует несколько приемов.

(1) Метод полярных координат.

Алгоритм P. *(Метод полярных координат для нормального распределения.)* Этот алгоритм вычисляет две независимые нормально распределенные случайные величины X_1 и X_2 по двум заданным независимым случайным числам U_1 и U_2.

P1 [Получить случайные числа.] Выработать два независимых случайных числа U_1, U_2, равномерно распределенных между нулем и единицей. Установить $V_1 \leftarrow 2U_1 - 1$, $V_2 \leftarrow 2U_2 - 1$. (Теперь V_1 и V_2 равномерно распределены между -1 и $+1$. Для большинства машин в этот момент предпочтительней представить V_1 и V_2 в форме с плавающей точкой.)

P2 [Вычислить S.] Установить $S \leftarrow V_1^2 + V_2^2$.

P3 [$S \ge 1$?] Если $S \ge 1$, вернуться к шагу **P1**. (Шаги с **P1** по **P3** в среднем выполняются 1.27 раз при стандартном отклонении 0.587; см. упр. 6).

Р4 [Вычислить X_1, X_2.] Присвоить X_1, X_2 значения, вычисленные по формулам

$$X_1 = V_1\sqrt{\frac{-2\ln S}{S}}, \qquad X_2 = V_2\sqrt{\frac{-2\ln S}{S}}. \tag{11}$$

Это и есть требующиеся значения нормально распределенных случайных величин. ∎

Чтобы доказать, что метод точен, воспользуемся элементарной аналитической геометрией. Если на шаге **Р3** $S < 1$, точка плоскости с декартовыми координатами (V_1, V_2) является *случайной точкой, равномерно распределенной внутри единичного круга.* Переходя к полярным координатам $V_1 = R\cos\Theta$, $V_2 = R\sin\Theta$, находим $S = R^2$, $X_1 = \sqrt{-2\ln S}\cos\Theta$, $X_2 = \sqrt{-2\ln S}\sin\Theta$. Используя также полярные координаты $X_1 = R'\cos\Theta'$, $X_2 = R'\sin\Theta'$, мы видим, что $\theta' = \theta$ и $R' = \sqrt{-2\ln S}$. Ясно, что R' и Θ' независимы, так как R и Θ независимы внутри единичного круга. Кроме того, Θ' равномерно распределена между 0 и 2π, а вероятность того, что $R' \le r$, равна вероятности события—$2\ln S \le r^2$, т. е. вероятности события $S \ge e^{-r^2/2}$. Последняя равна $1 - e^{-r^2/2}$, поскольку $S = R^2$ равномерно распределена между нулем и единицей. Вероятность того, что R' лежит между r и $r + dr$, равна поэтому производной от $1-e^{-r^2/2}$, а именно $re^{-r^2/2}dr$. Подобным же образом вероятность попадания Θ' в интервал между θ и $\theta + d\theta$ есть $(1/2\pi)d\theta$. Поэтому вероятность того, что $X_1 \le x_1$, а $X_2 \le x_2$, равна

$$\begin{aligned}\int_{\{\,(r,\theta)\mid r\cos\theta\le x_1,\, r\sin\theta\le x_2\,\}} \frac{1}{2\pi}e^{-r^2/2}r\,dr\,d\theta &= \\ &= \frac{1}{2\pi}\int_{\{\,(x,y)\mid x\le x_1, y\le x_2\,\}} e^{-(x^2+y^2)/2}\,dx\,dy = \\ &= \left(\sqrt{\frac{1}{2\pi}}\int_{-\infty}^{x_1} e^{-x^2/2}\,dx\right)\left(\sqrt{\frac{1}{2\pi}}\int_{-\infty}^{x_2} e^{-y^2/2}\,dy\right).\end{aligned}$$

Это доказывает, что X_1 и X_2 независимы и нормально распределены, как и требовалось. Алгоритм Р предложили Дж. Бокс, М. Маллер и Дж. Марсалья.

(2) Метод Марсальи (метод прямоугольника—клина—хвоста).

В этом методе мы используем распределение

$$F(x) = \sqrt{\frac{2}{\pi}}\int_0^x e^{-t^2/2}\,dt \qquad x \ge 0, \tag{12}$$

так что $F(x)$ является функцией распределения *модуля* нормальной случайной величины. После вычисления X в соответствии с этим распределением ее значению приписывается случайный знак, в результате мы получаем правильное значение нормальной случайной величины.

Общий подход Марсальи, позволяющий эффективно использовать случайные числа, будет проиллюстрирован на примере распределения **(12)**. Изучая этот частный пример, мы одновременно узнаем несколько общих методов.

За основу принимается следующее представление:

$$F(x) = p_1F_1(x) + p_2F_2(x) + \cdots + p_nF_n(x), \tag{13}$$

где F_1, F_2, ..., F_n—некоторые распределения. Случайная величина X тогда определяется по распределению F_j с вероятностью p_j. Некоторые из распределений $F_j(x)$ могут оказаться довольно трудными для вычислений, но мы обычно подбираем их так, чтобы в этом случае вероятность p_j была очень мала. Большинство функций распределения $F_j(x)$ окажутся очень простыми, будучи тривиальными модификациями равномерного распределения. Метод в целом позволяет создавать чрезвычайно эффективные программы, так как *среднее* время вычислений очень мало.

Метод легче понять, если иметь дело с *производными* функций распределения вместо самих распределений. Обозначим

$$f(x) = F'(x), \qquad f_j(x) = F_j'(x).$$

Функции $f(x)$ и $f_j(x)$ называются *плотностями* этих вероятностных распределений. Уравнение **(13)** преобразуется в

$$f(x) = p_1f_1(x) + p_2f_2(x) + \cdots + p_nf_n(x). \tag{14}$$

Каждая $f_j(x) \ge 0$, а полная площадь под кривой $f_j(x)$ равна 1. Поэтому существует простая графическая интерпретация соотношения **(14)**. Площадь под графиком $f(x)$ делится на n частей, каждая из

которых имеет площадь p_j и соответствует $f_j(x)$. Посмотрите на рис. 9, проясняющий ситуацию для случая, которым мы сейчас интересуемся, когда $f(x) = F'(x) = \sqrt{(2/\pi)}e^{-x^2/2}$. Площадь под кривой на рисунке разделена на $n = 37$ частей: 12 довольно

Picture: **Рис. 9. Частотная функция, состоящая из 37 частей. Случайная величина вычисляется в среднем столько раз какова площадь части, соответствующей ее частоте.**

больших прямоугольников, которые соответствуют $p_1 f_1(x), \ldots, p_{12} f_{12}(x)$, 12 узких небольших прямоугольников, примыкающих сверху к предыдущим и соответствующих $p_{13} f_{13}(x), \ldots, p_{24} f_{24}(x)$, 12 клинообразных фигур, представляющих функции $p_{25} f_{25}(x), \ldots, p_{36} f_{36}(x)$, и оставшаяся часть $p_{37} f_{37}(x)$, совпадающая с $f(x)$ при $x \geq 3$.

Ступеньки $f_1(x), \ldots, f_{12}(x)$ описывают *равномерные распределения.* Например, $f_3(x)$ отвечает случайной величине, равно- мерно распределенной между $1/2$ и $3/4$. Мы хотим определить $p_1, p_2, \ldots, p_{12}$ так, чтобы эффективно вырабатывать нормальные случайные величины. (Предположим, что мы имеем дело с двоичной машиной; для десятичной применяется аналогичная процедура.) Выберем их кратными, скажем, $1/256$. Это означает, что площадь под $p_j f_j(x)$ будет кратна $1/256$. Примем высоту $f_j(x)$ равной

$$\lfloor 64 f(j/4) \rfloor / 64 \qquad 1 \leq j \leq 12.$$

В таблице приводятся получающиеся при этом значения вероятностей:

$$\begin{array}{cccccccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} & p_{11} & p_{12} & (p_{13} + \cdots + p_{37}) \\ \frac{49}{256} & \frac{45}{256} & \frac{38}{256} & \frac{30}{256} & \frac{23}{256} & \frac{16}{256} & \frac{11}{256} & \frac{6}{256} & \frac{4}{256} & \frac{2}{256} & \frac{1}{256} & \frac{0}{256} & \frac{31}{256.} \end{array} \tag{15}$$

Отсюда следует, что реализация равномерных распределений $F_1, \ldots, F_{12}$ займет $p_1 + p_2 + \cdots + p_{12} = 88\%$ времени (88% площади под $f(x)$ занимают большие прямоугольники).

Марсальей был предложен эффективный способ выбора из 13 возможностей, представленных в (15). При заданном случайном двоичном целом числе в промежутке от 0 до 255, двоичное представление которого есть $b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8$, мы поступаем следующим образом:

$$\begin{array}{ll} \text{если} \quad 0 \leq b_1 b_2 b_3 b_4 < 10, & \text{полагаем } X = A[b_1 b_2 b_3 b_4] + \frac{1}{4}U; \\ \text{если} \quad 40 \leq b_1 b_2 b_3 b_4 b_5 b_6 < 52, & \text{полагаем } X = B[b_1 b_2 b_3 b_4 b_5 b_6] + \frac{1}{4}U; \\ \text{если} \quad 208 \leq b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8 < 225, & \text{полагаем } X = C[b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8] + \frac{1}{4}U; \\ \text{если} \quad 225 \leq b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8, & \text{вырабатываем } X, \text{используя распределения } F_{13}, \ldots, \\ & F_{37}, \text{ как описывается ниже.} \end{array} \tag{16}$$

Здесь A, B, C—вспомогательные массивы, приведенные в табл. 1, а U—случайное число (от 0 до 1).

Чтобы понять, почему эта процедура работает, рассмотрим, например, случай $j = 4$. Функция $F_4(x)$—это равномерное распределение случайной величины (между $3/4$ и 1), которое имеет случайная величина $X = \frac{3}{4} + \frac{1}{4}U$. Вероятность того, что это распределение реализуется в описанном выше методе, такова: $\frac{1}{16} \times$ (число раз, которое "$\frac{3}{4}$"встречается в массиве A) $+ \frac{1}{64} \times$ (аналогичное число "появлений" в массиве B) $+ \frac{1}{256} \times$ (число "появлений" в массиве C). Она равна $\frac{1}{16} + \frac{3}{64} + \frac{2}{256} = \frac{30}{256} = p_4$, так что $F_4(x)$ моделируется с правильной вероятностью. Конечно, было бы еще быстрее воспользоваться другим методом:

$$\text{"если } 0 \leq b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8 < 225, \text{ полагаем}$$
$$X = T[b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8] + \frac{1}{4}U\text{"}$$

вместо первых трех проверок (16). Здесь T задавалось бы таблицей из 225 чисел, 49 из которых равнялись бы "0", 45 чисел равнялись бы $\frac{1}{4}$, 38 раз повторялось бы $\frac{1}{2}$ и т. д. Для сравнения в предыдущем методе используется только таблица, содержащая 39 величин, и он ненамного медленней, так как проверка "$b_1 b_2 b_3 b_4 b_5 b_6 < 52$?", занимает не более $3/8$ всего времени, а проверка "$b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8 < 225$?" требует только $3/16$ времени. Экономия памяти в методе, подобном (16), еще значительней в общем случае, когда имеется больше 256 возможностей.

Таблица 1

Таблицы для эффективного выбора из 12 альтернатив в методе (16)

$$
\begin{array}{lll}
A[0]=0 & B[40]=\frac{1}{4} & C[208]=0 \\
A[1]=0 & B[41]=\frac{1}{4} & C[209]=\frac{1}{4} \\
A[2]=0 & B[42]=\frac{1}{4} & C[210]=\frac{1}{2} \\
A[3]=\frac{1}{4} & B[43]=\frac{1}{2} & C[211]=\frac{1}{2} \\
A[4]=\frac{1}{4} & B[44]=\frac{3}{4} & C[212]=\frac{3}{4} \\
A[5]=\frac{1}{2} & B[45]=\frac{3}{4} & C[213]=\frac{3}{4} \\
A[6]=\frac{1}{2} & B[46]=\frac{3}{4} & C[214]=1 \\
A[7]=\frac{3}{4} & B[47]=1 & C[215]=1 \\
A[8]=1 & B[48]=\frac{3}{2} & C[216]=1 \\
A[9]=\frac{5}{4} & B[49]=\frac{3}{2} & C[217]=\frac{3}{2} \\
 & B[50]=\frac{7}{4} & C[218]=\frac{3}{2} \\
 & B[51]=2 & C[219]=\frac{3}{2} \\
 & & C[220]=\frac{7}{4} \\
 & & C[221]=\frac{7}{4} \\
 & & C[222]=\frac{9}{4} \\
 & & C[223]=\frac{9}{4} \\
 & & C[224]=\frac{5}{2}
\end{array}
$$

Читателю следует в этом месте сделать паузу, чтобы тщательно изучить весь метод, пока он не поймет, из-за чего по формулам (16) распределения F_j выбираются с вероятностями p_j для $1 \le j \le 12$.

Распределения $F_{13}(x)$, ..., $F_{24}(x)$—также равномерные, и с ними можно поступать аналогично. Они имеют довольно малую вероятность $\left(\text{некоторое число между } 0 \text{ и } \frac{1}{256}\right)$, поэтому наиболее эффективно проверять эти вероятности с помощью части нашего алгоритма для выработки нормальных случайных величин, который будет использоваться не слишком часто (см. шаг M4 в алгоритме M, ниже). Эти распределения представляют собой поправки к большим прямоугольникам, возникающие при округлении вероятностей для (15). Мы имеем

$$p_j + p_{j+12} = \frac{1}{4} f(j/4), \qquad 1 \le j \le 12. \tag{12}$$

Теперь вернемся к вопросу о вычислении случайных величин для клинообразных распределений $F_{25}(x)$, ..., $F_{36}(x)$. Типичный случай изображен на рис. 10. Когда $x < 1$, кривая выпукла

Picture: Рис. 10. Функции плотности вероятности, для которых при выработке случайных чисел может быть использован алгоритм L.

вверх, а при $x > 1$—вниз, но в обоих случаях кривая очень близка к прямой линии и может быть вложена, как показано на рисунке, между двумя прямыми.

Алгоритм L. *(Почти линейные плотности.)* Этот алгоритм можно использовать для выработки значения случайной величины X для любой плотности распределения $f(x)$, удовлетворяющей следующим условиям (ср. с рис. 10):

$$\begin{gathered} f(x) = 0 \qquad \text{для } x < s \text{ и } x > s + h; \\ a - b(x-s)/h \le f(x) \le b - b(x-s)/h \qquad \text{для } s \le x \le s + h. \end{gathered} \tag{18}$$

L1 **[Получить $U \le V$.]** Выработать два независимых случайных числа U, V, равномерно распределенных между нулем и единицей. Если $U > V$, поменять местами $U \leftrightarrow V$.

L2 **[Простой случай?]** Если $V \le a/b$, перейти к **L4**.

L3 **[Попытаться еще раз?]** Если $V > U + (1/b)f(s + hU)$, вернуться обратно к шагу **L1**. (Если a/b близко к 1, этот шаг алгоритма будет использоваться не слишком часто.)

L4 **[Вычислить X.]** Установить $X \leftarrow s + hU$. ∎

Для доказательства правильности алгоритма заметим, что, когда мы приходим к шагу **L4**, точка (U, V)—это случайная

Picture: Рис. 11. Область "принятия результата" в алгоритме L.

точка в квадрате, изображенном на рис. 11, а именно $0 \le U \le V \le U + (1/b)f(s + hU)$.

Условия (18) гарантируют, что

$$\frac{a}{b} \le U + \frac{1}{b} f(s + hU) \le 1.$$

Теперь вероятность того, что $X \le s + hx$ для $0 \le x \le 1$, равна отнешению площади слева от вертикальной линии $U = x$ на рис. 11 ко всей площади, т. е.

$$\int_0^x \frac{1}{b} f(s + hu)\, du \Big/ \int_0^1 \frac{1}{b} f(s + hu)\, du = \int_s^{s+hx} f(v)\, dv;$$

поэтому X имеет нужное распределение.

Чтобы использовать этот алгоритм, нам необходимо определить a_j, b_j, s_j, h для плотностей вероятностей $f_{j+24}(x)$ (рис. 9). Нетрудно видеть, что при $1 \le j \le 12$

$$\begin{aligned} f_{j+24}(x) &= \frac{1}{p_j + 24}\sqrt{\frac{2}{\pi}}(e^{-x^2/2} - e^{-(j/4)^2/2}), \qquad s_j \le x \le s_j + h; \\ h &= \frac{1}{4}; s_j = (j-1)/4; \\ p_{j+24} &= \sqrt{\frac{2}{\pi}} \int_{s_j}^{s_j+h} (e^{-t^2/2} - e^{-(j/4)^2/2})\, dt. \end{aligned} \tag{19}$$

Кроме того,

$$\begin{aligned} a_j &= f_{j+24}(s_j) && \text{при } 1 \le j \le 4, \\ b_j &= f_{j+24}(s_j) && \text{при } 5 \le j \le 12; \\ b_j &= -h f'_{j+24}(s_j + h) && \text{при } 1 \le j \le 4, \\ a_j &= f_{j+24}(x_j) + (x_j - s_j) b_j / h && \text{при } 5 \le j \le 12, \end{aligned} \tag{20}$$

где x_j—корень уравнения $f'_{j+24}(x_j) = -b_j/h$.

Последнее распределение $F_{37}(x)$ должно моделироваться только один раз из четырехсот. Оно используется всегда, когда должен

Picture: Рис. 12. Алгоритм "прямоугольник-клин-хвост" для выработки нормально распределенных случайных величин.

получиться результат $X \ge 3$. В этом случае можно применить модификацию алгоритма P, как показано ниже (шаги M8—M9). Предоставляем читателю возможность доказать, что приведенный метод точен.

Теперь приводим процедуру полностью.

Алгоритм M. *(Метод Марсальи—Макларена для нормальных случайных величин.)* В этом алгоритме используются некоторые вспомогательные таблицы, устроенные, как объяснялось в тексте (примеры приведены в табл. 1 и 2). Алгоритм приводится для двоичной машины, для десятичной

машины он строится аналогично.

Таблица 2[10]

Примеры таблиц, используемых в алгоритме M

j	$S[j]$	$P[j]$	$Q[j]$	$D[j]$	$E[j]$
1	0	0.885	0.881	0.51	16
2	$\frac{1}{4}$	0.895	0.885	0.79	8
3	$\frac{1}{2}$	0.910	0.897	0.90	5.33
4	$\frac{3}{4}$	0.929	0.914	0.98	4
5	1	0.945	0.930	0.99	3.08
6	$\frac{5}{4}$	0.960	0.947	0.99	2.44
7	$\frac{3}{2}$	0.971	0.960	0.98	2.00
8	$\frac{7}{4}$	0.982	0.974	0.96	1.67
9	2	0.987	0.982	0.95	1.43
10	$\frac{9}{4}$	0.991	0.989	0.93	1.23
11	$\frac{5}{2}$	0.994	0.992	0.94	1.08
12	$\frac{11}{4}$	0.997	0.996	0.94	0.95
13	3	1.000			

M1 **[Получить U.]** Выработать случайное число $U = .b_0b_1b_2\ldots b_t$. (Здесь b—биты в двоичном представлении U. Для хорошей точности t должно быть не меньше 24.) Установить $\psi \leftarrow b_0$. (Позже ψ понадобится для определения знака результата.)

M2 **[Большой прямоугольник?]** Если $b_1b_2b_3b_4 < 10$, где "$b_1b_2b_3b_4$" обозначает двоичное целое число $8b_1 + 4b_2 + 2b_3 + b_4$, установить

$$X \leftarrow A[b_1b_2b_3b_4] + .00b_5b_6\ldots b_t$$

и перейти к **M10**. Иначе, если $b_1b_2b_3b_4b_5b_6 < 52$, установить

$$X \leftarrow B[b_1b_2b_3b_4b_5b_6] + .00b_7b_8\ldots b_t$$

и перейти к **M10**. Иначе, если $b_1b_2b_3b_4b_5b_6b_7b_8 < 225$, установить

$$X \leftarrow C[b_1b_2b_3b_4b_5b_6b_7b_8] + .00b_9b_{10}\ldots b_t$$

и перейти к **M10**.

M3 **[Клин или хвост?]** Найти *наименьшее* значение j, $1 \le j \le 13$, для которого $b_1b_2\ldots b_t < P[j]$. Если $j = 13$, перейти к шагу **M8**.

M4 **[Узкий прямоугольник?]** Если $.b_1b_2\ldots b_t < Q[j]$, выработать новое случайное число U, установить $X \leftarrow S[j] + \frac{1}{4}U$ и перейти к **M10**.

M5 **[Получить $U \le V$.]** Выработать два новых случайных числа U, V; если $U > V$, поменять их местами $U \leftrightarrow V$. (Теперь выполняется алгоритм L.) Установить $X \leftarrow S[j] + \frac{1}{4}U$.

M6 **[Простой случай?]** Если $V \le D[j]$, перейти к **M10**.

M7 **[Еще одна попытка?]** Если $V > U + E[j](e^{-(X^2-S[j+1]^2)/2} - 1)$, вернуться к шагу **M5**; в противном случае перейти к **M10**. (Этот шаг выполняется с малой вероятностью.)

M8 **[Получить $U^2 + V^2 < 1$.]** Выработать два новых случайных числа U, V. Установить $W \leftarrow U^2 + V^2$. Если $W \ge 1$, повторить этот шаг.

M9 **[Вычислить $X \ge 3$.]** Установить $T \leftarrow \sqrt{(9 - 2\ln W)/W}$. Установить $X \leftarrow U \times T$. Если $X > 3$, перейти к **M10**; в противном случае установить $X \leftarrow V \times T$. Если $X \ge 3$, перейти к **M10**; иначе вернуться к шагу **M8**. (Последнее происходит в половине всех случаев, когда выполняется данный шаг.)

M10 **[Присвоить знак.]** Если $\psi = 1$, установить $X \leftarrow -X$. ∎

Весь алгоритм являет собой весьма приятный пример математической теории, густо сдобренной изобретательностью программиста. Это прекрасная иллюстрация искусства программирования.

---

[10] На практике данные для таблиц P, Q, D, E следует давать с большей точностью.

Таблицы A, B и C уже были описаны. Остальные таблицы, необходимые для алгоритма M, строятся следующим образом;

$$\begin{aligned}
&S[j] = (j-1)/4, \qquad 1 \le j \le 13;\\
&P[j] = p_1 + p_2 + \cdots + p_{12} + (p_{13} + p_{25}) + \cdots + (p_{12+j} + p_{24+j}), \qquad 1 \le j \le 12;\ P[13] = 1;\\
&Q[j] = P[j] - p_{24+j}, \qquad 1 \le j \le 12;\\
&D[j] = a_j/b_j, \qquad 1 \le j \le 12;\\
&E[j] = \sqrt{\frac{2}{\pi}}e^{-(j/4)^2/2}/b_j p_{j+24}, \qquad 1 \le j \le 12.
\end{aligned} \tag{21}$$

[Величины a_j, b_j, p_{j+24} определяются в (19) и (20).]

В табл. 2 значения приводятся только с несколькими значащими цифрами, но в настоящей программе они должны иметь точность, соответствующую полному машинному слову. Для всех вспомогательных таблиц алгоритма M требуется 101 машинное слово.

Этот метод чрезвычайно быстрый, так как 88% времени работают только шаги M1, M2 и M10, остальные же шаги также не слишком медленные. На рис. 9 мы разделили интервал от 0 до 3 на 12 частей. Если бы мы разделили его на большее число частей, скажем 48, понадобились бы более длинные таблицы, но зато при этом в 97% случаев вычисления ограничивались только шагами M1, M2, M10. Полные таблицы как для двоичных, так и для десятичных машин приводятся в статье Марсальи, Макларена и Брэя (*CACM,* **7** (1964), 4–10). Там для экономии памяти разработан дополнительный прием, связанный с перекрыванием частей таблиц A, B, C и S.)

(3) Метод Тейчроева. Нормальные случайные величины можно получить также следующим образом. Выработаем 12 независимых случайных чисел $U_1, U_2, \ldots, U_{12}$, равномерно распределенных между нулем и единицей. Положим $R = (U_1 + U_2 + \cdots + U_{12} - 6)/4$. Вычислим

$$X = ((((a_9R^2 + a_7)R^2 + a_5)R^2 + a_3)R^2 + a_1)R, \tag{22}$$

где

$$\begin{gathered}
a_1 = 3.94984\,6138, \quad a_3 = 0.25240\,8784,\\
a_5 = 0.07654\,2912, \quad a_7 = 0.00835\,5968, \quad a_9 = 0.02989\,9776.
\end{gathered} \tag{23}$$

Такое X будет хорошим приближением для нормальной случайной величины. Никогда не получается слишком больших значений X, но с вероятностью, меньшей $1/50000$, вырабатываются значения, превышающие те, где метод работает правильно.

Метод основан на том, что R имеет *приблизительно* нормальное распределение со средним значением нуль и стандартным отклонением $\frac{1}{4}$. Пусть $F_1(x)$—истинное распределение для R, а $F(x)$—нормальное распределение, определяемое формулой (10). Положим $X = F^{-1}(F_1(R))$; так как $F_1(R)$—равномерно распределенная случайная величина, X будет распределена нормально. Формула (22) представляет приближение функции $F^{-1}(F_1(R))$ полиномом в промежутке $|R| \le 1$.

(4) Сравнение методов. Мы привели три метода для выработки нормальных случайных величин. Метод полярных координат довольно медленный, но обеспечивает абсолютную точность. Его легко запрограммировать, если есть стандартные программы для вычисления квадратного корня и логарифма. Метод Тейчроева также легко программируется, для него не нужно других подпрограмм. Поэтому он нуждается в меньшей памяти. Метод этот приближенный, хотя для большинства приложений дает достаточную точность (ошибка не превышает 2×10^{-4} при $|R| \le 1$). Метод Марсальи значительно быстрее любых других и подобно методу полярных координат имеет абсолютную точность. Для него необходимы подпрограммы квадратного корня, логарифма и показательной функции и, кроме того, вспомогательные таблицы для 100–400 констант. Поэтому требования к памяти довольно высокие. Однако на больших машинах скорость метода с избытком компенсирует этот недостаток. Программу для метода Марсальи написать гораздо труднее, но если подпрограмму, основанную на алгоритме M, составить в общем виде, она явится ценным вкладом в любую библиотеку подпрограмм. Многочисленные применения нормально распределенных случайных величин требуют большого количества случайных чисел, так что важна скорость их выработки.

Дополнительную информацию о методе Тейчроева, а также обзор некоторых других методов, худших, как теперь выяснилось, чем обсуждаемые здесь, можно получить из статьи М. Маллера (*JACM,* **6** (1959), 376–383).

(5) Разновидности нормального распределения. Мы рассмотрели нормальное распределение с нулевым средним значением и стандартным отклонением, равным единице. Если X имеет такое распределение, то у функции распределения случайной величины

$$Y = \mu + \sigma X \tag{24}$$

среднее значение равно μ, а стандартное отклонение σ. Более того, если X_1 и X_2—независимые нормальные случайные величины со средним значением нуль и единичным стандартным отклонением и если

$$Y_1 = \mu_1 + \sigma_1 X_1, \qquad Y_2 = \mu_2 + \sigma_2(\rho X_1 + \sqrt{1-\rho^2}X_2), \tag{25}$$

то Y_1 и Y_2—*зависимые* случайные величины, распределенные со средними значениями μ_1, μ_2, стандартными отклонениями σ_1, σ_2 и коэффициентом корреляции ρ. (Обобщение на случай n переменных см. в упр. 13.)

D. Экспоненциальное распределение. Другой важный вид случайных величин—величины с *экспоненциальным распределением*. Такие случайные величины бывают нужны в задачах, где рассматривается "время появления". Например, если радиоактивное вещество излучает в среднем каждые μ секунд одну альфа-частицу, то промежутки времени между двумя последовательными вылетами частиц имеют экспоненциальное распределение со средним значением μ. Это распределение определяется формулой

$$F(x) = 1 - e^{-x/\mu}, \qquad x \geq 0. \tag{26}$$

Отсюда следует, что если X имеет экспоненциальное распределение со средним значением 1, то μX подчиняется экспоненциальному распределению со средним μ. Поэтому достаточно рассмотреть случай $\mu = 1$. Обычно используются три метода.

(1) Логарифмический метод. Ясно, что $y = F(x) = 1 - e^{-x}$ можно представить в виде $x = F^{-1}(y) = -\ln(1-y)$. Поэтому, вследствие соотношения (6), величина $-\ln(1-U)$ имеет экспоненциальное распределение. Так как $1-U$ распределена равномерно, если U—равномерно распределенное случайное число, то случайная величина

$$X = -\ln U \tag{27}$$

распределена экспоненциально со средним значением, равным единице. (В программах следует избегать случая $U = 0$.)

(2) Метод случайной минимизации. Следующий алгоритм (Дж. Марсалья) вычисляет значения экспоненциально распределенной случайной величины без использования подпрограммы логарифма.

Алгоритм E. *(Экспоненциальное распределение со средним* 1*.)* Используются таблицы констант $P[j]$, $Q[j]$ для $j \geq 1$, определенные формулами

$$P[j] = 1 - \frac{1}{e^j}, \quad Q[j] = \frac{1}{e-1}\left(\frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{j!}\right). \tag{28}$$

Длина таблиц ограничивается значением максимальной дроби, которую можно разместить в машинном слове.

E1 [Начало дробной части.] Установить $j \leftarrow 1$. Выработать случайные числа U_0 и U_1 и установить $X \leftarrow -U_1$.

E2 [Минимизация закончена?] Если $U_0 < Q[j]$, перейти к **E4**.

E3 [Минимизировать.] Установить $j \leftarrow j+1$. Выработать случайное число U_j; если $X > U_j$, установить $X \leftarrow U_j$. Вернуться обратно к шагу **E2**.

E4 [Начало целой части.] (Мы уже вычислили дробную часть окончательного результата, X, и должны добавить к нему соответствующее целое число, чтобы закончить вычисления.) Выработать новое случайное число U и установить $j \leftarrow 1$.

E5 [Сделана ли поправка?] Если $U < P[j]$, алгоритм заканчивается.

E6 [Поправка на 1.] Установить $j \leftarrow j+1$, $X \leftarrow X+1$ и вернуться к шагу **E5**. ∎

Чтобы показать справедливость метода, проанализируем распределение X в начале шага E4. Если n—окончательное значение j, мы имеем $X = \min(U_1, U_2, \ldots, U_n)$, где U_1, U_2, $\ldots$, U_n — независимые случайные числа; поэтому вероятность того, что $X \leq x$, есть $p_n(x) = 1 - (1-x)^n$. Вероятность того, что n—окончательное значение, равна $Q[n] - Q[n-1] = 1/(e-1)n!$. Поэтому полная вероятность события $X \leq x$ равна

$$\sum_{n\geq 1} \frac{p(x)}{(e-1)n!} = \frac{e}{e-1}(1-e^{-x}), \qquad 0 \leq x \leq 1.$$

Аналогично, при рассмотрении шагов E4–E6 мы находим, что $\lfloor X \rfloor \leq m$ с вероятностью $P[m+1]$. Окончательно, вероятность того, что $m \leq X \leq m+x$, есть

$$(P[m+1] - P[m])\left(\frac{e}{e-1}(1-e^{-x}\right) = e^{-m} - e^{-(m+x)}, \qquad 0 \leq x \leq 1.$$

Это доказывает, что X имеет распределение $F(x) = 1 - e^{-x}$ для $0 \le x < \infty$.

Этот алгоритм довольно быстрый. В среднем шаги **E2** и **E5** выполняются только 1.582 раз и только 2.582 случайных числа вычисляется для получения одной экспоненциальной случайной величины. Это может оказаться значительно быстрее, чем вычисление логарифма одного случайного числа. Интересная модификация метода содержится в работе [M. Sibuya, *Ann. Inst. Stat. Math.*, **13** (1962), 231–237].

(3) Метод прямоугольника—клина—хвоста. Для нормального распределения (как и для многих других) существует очень быстрый метод, основанный на разложении функции плотности вероятности. Детали обсуждаются в *CACM*, **7** (1964), 298–300.

E. Другие непрерывные распределения. Мы перечислим здесь коротко некоторые другие распределения, которые бывают довольно часто нужны на практике, и методы моделирования соответствующих случайных величин с помощью уже известных нам приемов.

χ^2-распределение с ν степенями свободы, которое также называют гамма-распределением порядка $\nu/2$. Мы имеем

$$F(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^x t^{\nu/2-1} e^{-t/2}\, dt, \qquad x \ge 0. \tag{29}$$

Если $\nu = 2k$, где k—целое число, положим $X = 2(Y_1 + Y_2 + \cdots + Y_k)$, где Y_i—независимые экспоненциально распределенные случайные величины, со средним значением 1 каждая. Если $\nu = 2k + 1$, положим $X = 2(Y_1 + \cdots + Y_k) + Z^2$, где Y_i—те же величины, а Z—независимая нормально распределенная случайная величина (среднее 0, дисперсия 1). Для доказательства см. упр. **16**. Заметим, что, если $Y_1, \ldots, Y_k$ находятся логарифмическим методом, для определения $Y_1 + \cdots + Y_k = -\ln(U_1 \ldots U_k)$ требуется вычислить только один логарифм.

(2) Бета-распределение с ν_1 и ν_2 степенями свободы, определяется формулой:

$$F(x) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \int_0^x t^{\nu_1/2-1}(1-t)^{\nu_2/2-1}\, dt, \qquad 0 \le x \le 1. \tag{30}$$

Пусть Y_1 и Y_2 независимы и имеют χ^2-распределение с ν_1, ν_2 степенями свободы соответственно. Полагаем $X \leftarrow Y_1/(Y_1 + Y_2)$.

Другой метод, который остается справедливым и для нецелых ν_1 и ν_2, сводится к вычислению $Y_1 \leftarrow U_1^{2/\nu_1}$, $Y_2 \leftarrow U_2^{2/\nu_2}$ с повторением в случае необходимости этого процесса до тех пор, пока не будет выполнено неравенство $Y_1 + Y_2 \le 1$. Наконец, $X \leftarrow Y_1/(Y_1 + Y_2)$. [См. M. D. Jöhnk, *Metrika*, **8** (1964), 5–15.] По-другому, если $\nu_1 = 2k_1$, а $\nu_2 = 2k_2$—оба четные целые числа, мы можем определить X как k_1-е по порядку наименьшее из $k_1 + k_2 - 1$ независимых случайных чисел.

(3) F-распределение (распределение отношений дисперсий) с ν_1 и ν_2 степенями свободы определяется как

$$F(x) = \frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \int_0^x t^{\nu_1/2-1}(\nu_2+\nu_1 t)^{-\nu_1/2-\nu_2/2}\, dt, \qquad x \ge 0. \tag{31}$$

Пусть Y_1 и Y_2—независимые случайные величины, имеющие χ^2-распределение с ν_1, ν_2 степенями свободы соответственно. Положим $X \leftarrow Y_1\nu_2/Y_2\nu_1$, или $X \leftarrow \nu_2 Y/\nu_1(1-Y)$, где Y имеет бета-распределение **(30)**.

(4) t-распределение с ν степенями свободы определяется как

$$F(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)} \int_{-\infty}^x (1 + t^2/\nu)^{-(\nu+1)/2}\, dt. \tag{32}$$

Пусть Y_1 — нормально распределенная случайная величина (среднее 0, дисперсия 1), пусть Y_2 — независимая от Y_1 случайная величина, имеющая χ^2-распределение с ν степенями свободы. Положим $X \leftarrow Y_1/\sqrt{Y_2/\nu}$.

(5) Случайная точка на n-мерной сфере с единичным радиусом. Пусть $X_1, X_2, \ldots, X_n$ — независимые нормальные случайные величины (среднее 0, дисперсия 1). Случайная точка на единичной сфере имеет координаты

$$(X_1/r, X_2/r, \ldots, X_n/r), \qquad \text{где } r = \sqrt{X_1^2 + X_2^2 + \cdots + X_n^2}. \tag{33}$$

Заметим, что если X_i вычисляются с помощью метода полярных координат (алгоритм **P**), то каждый раз в результате мы имеем две независимые величины X_1 и X_2 и $X_1^2 + X_2^2 = -2\ln S$ (в обозначениях

алгоритма P). Это экономит немного времени, необходимого для вычисления r. Справедливость метода проистекает из того факта, что функция распределения точки $(X_1, \ldots, X_n)$ имеет плотность, зависящую только от расстояния до центра. Поэтому проекция такой точки на единичную сферу имеет равномерное распределение. Метод впервые был предложен Дж. Брауном.

F. Важные дискретные распределения. Вообще говоря, к вероятностным распределениям случайных величин, принимающих только целочисленные значения, применимы приемы, описанные в начале этого раздела. Но некоторые из этих распределений столь важны для практических применений, что заслуживают специального рассмотрения.

(1) Геометрическое распределение. Если некоторое событие происходит с вероятностью p, число N независимых испытаний, необходимых, чтобы это событие произошло (или число испытаний между событиями), подчиняется геометрическому распределению. Мы имеем $N = 1$ с вероятностью p, $N = 2$ с вероятностью $(1-p)p$, ..., $N = n$ с вероятностью $(1-p)^{n-1}p$. (В сущности это та же ситуация, с которой мы встречались при "проверке интервалов" в п. 3.3.2; она возникает, когда мы интересуемся, сколько раз выполняются определенные циклы в алгоритмах этого раздела, например шаги P1–P3 в методе полярных координат.)

Для выработки значения случайной величины с таким распределением, когда p мало, есть удобная формула:

$$N = \lceil \ln U / \ln(1-p) \rceil. \tag{34}$$

Чтобы проверить ее, убедимся, что $\lceil \ln U/\ln(1-p)\rceil = n$ тогда и только тогда, когда $n-1 < \ln U/\ln(1-p) \le n$, т. е. $(1-p)^{n-1} > U \ge (1-p)^n$, а это происходит с вероятностью $p(1-p)^{n-1}$, что и требовалось показать.

Частный случай $p = 1/2$ еще легче моделировать на двоичной машине, так как формула (34) превращается в $N = \lceil -\log_2 U \rceil$, т. е. N на единицу больше, чем число первых нулевых разрядов в двоичном представлении U.

(2) Биномиальное распределение (t, p). Если некоторое событие происходит с вероятностью p, и мы проводим t независимых испытаний, полное число N происходящих при этом событий равно n с вероятностью $\binom{t}{n} p^n (1-p)^{t-n}$ (см. п. 1.2.10). Для этого распределения нет какого-либо прямого метода, аналогичного (34). Однако мы могли бы использовать то обстоятельство, что если N_1 имеет биномиальное распределение (t_1, p) и если, независимо, N_2 имеет биномиальное распределение (t_2, p), то $N_1 + N_2$ имеет биномиальное распределение $(t_1 + t_2, p)$. Когда t велико, биномиальное распределение приближенно описывается нормальным распределением со средним tp и среднеквадратичным отклонением $\sqrt{tp(1-p)}$. См. также прием, рассмотренный в упр. 25.

(3) Распределение Пуассона со средним значением μ. Это распределение так же связано с экспоненциальным распределением, как биномиальное с геометрическим. Оно характеризует число реализации в единицу времени событий, каждое из которых может произойти в любой момент. Например, число излучаемых в секунду альфа-частиц имеет распределение Пуассона. Вероятность того, что $N = n$, равна

$$e^{-\mu}\mu^n/n!, \qquad n \ge 0. \tag{35}$$

Если N_1, N_2—независимые пуассоновские случайные величины со средними μ_1, μ_2, то вероятность того, что $N_1 + N_2 = n$, равна

$$\sum_{0\le k\le n} \frac{e^{-\mu_1}\mu_1^k}{k!}\frac{e^{-\mu_2}\mu_2^{n-k}}{(n-k)!} = \frac{e^{-(\mu_1+\mu_2)}(\mu_1+\mu_2)^n}{n!}.$$

Таким образом, $N_1 + N_2$ имеет распределение Пуассона со средним значением $(\mu_1 + \mu_2)$.

Предположим, что мы хотим написать общую подпрограмму, вырабатывающую значения пуассоновских случайных величин со средним μ, где μ задается при входе в подпрограмму.

Алгоритм Q. *(Распределение Пуассона с произвольным μ.)*

Q1 [Вычислить экспоненту.] Присвоить $p \leftarrow e^{-\mu}$ и $N \leftarrow 0$, $q \leftarrow 1$. (Хотя $e^{-\mu}$ обычно вычисляется с помощью арифметики с плавающей точкой с привлечением стандартной подпрограммы, возможно, разумней пользоваться арифметикой с фиксированной точкой, правильно выбрав масштаб и округление для последующих операций с p и q.)

Q2 [Получить случайное число.] Выработать случайное число U, равномерно распределенное между 0 и 1.

Q3 [Умножить.] Установить $q \leftarrow qU$.

Q4 [Проверить, меньше ли $e^{-\mu}$.] Если $q \ge p$, установить $N \leftarrow N + 1$ и вернуться к шагу **Q2**. В противном случае алгоритм заканчивается выводом N. ∎

Чтобы доказать справедливость метода, заметим, что независимые равномерно распределенные случайные величины удовлетворяют условиям

$$U_1 \ge p, \quad U_1U_2 \ge p, \quad \ldots, \quad U_1U_2\ldots U_n \ge p, \quad U_1U_2\ldots U_{n+1} < p$$

с вероятностью

$$p\left(\ln\frac{1}{p}\right)^n \Big/ n! \qquad \text{для } 0 < p \le 1.$$

Применяя индукцию по n, посредством интегрирования

$$\int_p^1 (p/u_1)(\ln(u_1/p))^{n-1}\, du_1/(n-1)!$$

получим нужный результат.

Другой метод выработки случайных величин с распределением Пуассона, принадлежащий П. Крибс, основан на упоминавшемся ранее свойстве суммирования. При этом требуется более тщательное программирование, чем при реализации алгоритма Q, но при высоком уровне программирования его можно сделать более быстрым.

Алгоритм K. *(Распределение Пуассона с произвольным μ.)* Используется вспомогательная таблица $M[1] < M[2] < \cdots < M[n]$, описанная ниже[11].

K1 **[Начальная установка.]** Установить $m \leftarrow \mu$, $j \leftarrow n$, $N \leftarrow 0$.

K2 **[$m \ge M[j]$?]** Если $m < M[j]$, перейти к шагу **K5**.

K3 **[Выработать для среднего $M[j]$.]** Выработать значение X случайной целочисленной величины, имеющей распределение Пуассона со средним $M[j]$. (Это можно эффективно сделать, используя заранее составленные таблицы в соответствии с общим методом для целочисленных распределений, рассмотренным в упр. 21.)

K4 **[Изменить N, m.]** Установить $m \leftarrow m - M[j]$, $N \leftarrow N + X$ и вернуться к шагу **K2**.

K5 **[Уменьшить j.]** Уменьшить j на 1. Если $j > 0$, вернуться к шагу **K2**, в противном случае алгоритм заканчивается. ∎

Чтобы использовать этот алгоритм, мы должны составить специальные программы для частных значений μ, заданных в таблице $M[1]$, $M[2]$, $\ldots$, $M[n]$.

Например, мы могли бы принять $n = 10$, тогда

$$\begin{array}{rccccccccc} j = & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ M[j] = & 2^{-15} & 2^{-12} & 2^{-9} & 2^{-6} & 2^{-3} & 2^{-1} & 1 & 2 & 4 & 8 \end{array} \tag{36}$$

Этот метод неэффективен для больших значений μ, скажем $\mu \ge 50$. При $\mu < M[1] = 2^{-15}$ описанный алгоритм присваивает $N = 0$, так как вероятность того, что $N > 0$, равна $1 - e^{-\mu}$, т. е. меньше $\frac{1}{32\,000}$. Распределение Пуассона для малых значений μ моделировать чрезвычайно легко, так как для всех практических целей N будет довольно мало. Только для распределений с $M[j] = 4$ и 8 из приведенного выше списка значений потребуются большие таблицы при выполнении шага **K3**.

Для больших значений μ Аренс предложил эффективный, но довольно сложный метод порядка $\sqrt{\mu}$. Его процедура делит пуассоновское распределение на две части, одна из которых напоминает равнобедренный треугольник.

Упражнения

1. **[10]** Как вы предложите вырабатывать случайные числа, равномерно распределенные между случайными числами α и β ($\alpha < \beta$)?
2. **[M16]** Предполагая, что mU—случайное целое число между 0 и $m-1$, найдите *точную* вероятность того, что $\lfloor kU \rfloor = r$, если $0 \le r < k$. Сравните результат с требующейся вероятностью $1/k$.

>3. **[14]** Обсудите, что получится, если трактовать U как целое число и вырабатывать из него случайное целое между 0 и $k-1$, *деля* U на k вместо предложенного в тексте умножения. Таким образом, (1) следует изменить так:

```
01 ENTA  0
02 LDX   U
03 DIV   K
```

[11] В этом алгоритме n означает совсем не то, что в предыдущих абзацах и задается достаточно произвольно; см. ниже.—*Прим. перев.*

Результат окажется в регистре X. Хороший ли это метод?

4. **[M20]** Докажите оба соотношения в (7).

>5. **[21]** Предложите эффективный метод вычисления случайной величины с распределением $px + qx^2 + rx^3$, где $p \geq 0$, $q \geq 0$, $r \geq 0$ и $p + q + r = 1$.

>6. **[BM21]** Величина X вычисляется следующим методом.

"*Шаг 1.* Выработать два случайных числа U, V. *Шаг 2.* Если $U^2 + V^2 \geq 1$, вернуться к шагу 1, иначе установить $X \leftarrow U$."

Какова функция распределения X? Как часто будет выполняться шаг 1? (Определите среднее и среднеквадратичное отклонение.)

7. **[M18]** Объясните, почему в методе Марсальи для нормальных случайных величин желание выбрать p_j кратными $1/256$ приводит к формуле $p_j = \lfloor 64f(j/4)\rfloor/256$, $1 \leq j \leq 12$.

8. **[10]** Зачем особо выделять узкие прямоугольники $f_{13}, \ldots, f_{24}$ в методе Марсальи наравне с большими $f_1, \ldots, f_{12}$? (Почему это лучше, чем объединение каждой пары (f_1, f_{13}), (f_2, f_{14}), $\ldots$ в один большой прямоугольник?)

9. **[BM10]** Почему кривая $f(x)$ на рис. 9 выпукла вверх при $x < 1$ и вниз для $x > 1$?

10. **[BM21]** Выведите формулы для a_j, b_j в соотношении (20). Покажите также, что $E[j] = 16/j$, если $1 \leq j \leq 4$; $E[j] = 1/(e^{j/16-1/32} - 1)$, если $5 \leq j \leq 12$.

>11. **[BM27]** Докажите, что шаги M8–M9 в алгоритме M вырабатывают значение случайной величины, соответствующей хвосту нормального распределения, т. е. при $x \geq 3$ вероятность того, что $X < x$, равна

$$\int_3^x e^{-t^2/2}\,dt \Big/ \int_3^\infty e^{-t^2/2}\,dt.$$

12. **[BM46]** Полином Тейчроева (22)—это усеченная сумма чебышевских полиномов, так что не может быть лучшего приближения полиномом степени 9 к функции $F^{-1}(F_1(R))$. Можно ли использовать лучший полином?

13. **[BM25]** Задано множество n независимых нормальных случайных величин X_1, X_2, $\ldots$, X_n со средним 0 и дисперсией 1. Покажите, как найти константы b_j и a_{ij}, $1 \leq j \leq i \leq n$, такие, что если

$$Y_1 = b_1 + a_{11}X_1, \quad Y_2 = b_2 + a_{21}X_1 + a_{22}X_2, \quad \ldots, \quad Y_n = b_n + a_{n1}X_1 + a_{n2}X_2 + \cdots + a_{nn}X_n,$$

то Y_1, Y_2, $\ldots$, Y_n—зависимые, нормально распределенные случайные величины с заданной матрицей ковариаций (c_{ij}), и каждая случайная величина Y_j имеет среднее μ_j. (Ковариация c_{ij} величин Y_i и Y_j определяется как среднее значение $(Y_i - \mu_i)(Y_j - \mu_j)$. В частности, c_{jj} является дисперсией Y_j, квадратом среднеквадратичного отклонения. Не всякая матрица (c_{ij}) может быть матрицей ковариаций, и предполагается, конечно, что ваш метод должен работать лишь в том случае, если решение задачи существует.)

14. **[M20]** Если X—случайная величина с непрерывной функцией распределения $F(x)$, а c — константа, каково распределение cX?

15. **[BM21]** Если X_1 и X_2—независимые случайные величины с функциями распределения $F_1(x)$ и $F_2(x)$ и плотностями $f_1(x) = F_1'(x)$, $f_2(x) = F_2'(x)$, то каковы функции распределения и плотности вероятности случайной величины $X_1 + X_2$?

16. **[BM25]** (a) Покажите, что χ^2-распределение с одной степенью свободы—это распределение для X^2, где X—нормальная случайная величина со средним, равным нулю, и дисперсией, равной единице. (b) Покажите, что χ^2-распределение с двумя степенями свободы—это экспоненциальное распределение со средним 2. (c) Покажите, что если X_1 и X_2—независимые случайные величины, имеющие χ^2-распределение с ν_1 и ν_2 степенями свободы, то $X_1 + X_2$ имеет χ^2-распределение с $\nu_1 + \nu_2$ степенями свободы.

>17. **[M24]** Какова *функция распределения* $F(x)$ для геометрического распределения с вероятностью p? Какова *производящая функция* $G(z)$? Чему равны среднее и среднеквадратичное отклонение этого распределения?

18. **[M24]** Предложите метод вычисления случайного целого числа N, такого, что оно принимает значение n с вероятностью $np^2(1-p)^{n-1}$, $n \geq 0$. (Особый интерес представляет случай, когда p очень мало.)

19. **[M22]** (a) Найдите, сколько раз в среднем выполняется шаг E2 в алгоритме E; чему равно среднеквадратичное отклонение? (b) Тот же вопрос для шага E6.

>20. **[23]** Предположим, что мы хотим вычислить случайное значение величины X с таким дискретным распределением:

Значение $X =$	x_1	x_2	x_3	x_4	x_5	x_6
Вероятность $=$	$\frac{90}{512}$	$\frac{81}{512}$	$\frac{131}{512}$	$\frac{10}{512}$	$\frac{32}{512}$	$\frac{168}{512}$

Покажите, как можно эффективно вычислять X, используя девятиразрядное двоичное случайное число $b_1b_2b_3b_4b_5b_6b_7b_8b_9$, определенное аналогично формуле (16) с помощью вспомогательных таблиц, содержащих не более 35 элементов.

>21. [24] Ситуация в предыдущем упражнении несколько идеализирована, потому что обычно так не бывает, чтобы все вероятности имели один и тот же столь удобный множитель, как $1/512$. Предположим, что вероятности в предыдущем упражнении были изменены, а в действительности они должны быть такими:

$$\begin{array}{rcccccc} \text{Значение } X = & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ \text{Вероятность} = & \frac{89}{512}+\varepsilon_1 & \frac{80}{512}+\varepsilon_2 & \frac{131}{512}+\varepsilon_3 & \frac{10}{512}+\varepsilon_4 & \frac{32}{512}+\varepsilon_5 & \frac{168}{512}+\varepsilon_6 \end{array}$$

Здесь $0 \le +\varepsilon_j < \frac{1}{512}$ и $+\varepsilon_1++\varepsilon_2+\varepsilon_3+\varepsilon_4+\varepsilon_5+\varepsilon_6 = \frac{2}{512}$. Покажите, как эффективно выбрать эти числа для данного равномерно распределенного случайного числа $U = b_1b_2\ldots b_t$, где t сравнительно велико. Как и в упр. 20, попытайтесь использовать вспомогательные таблицы, содержащие не более 35 элементов.

22. [BM46] Можно ли получить точное пуассоновское распределение для больших μ, выработав соответствующую нормальную случайную величину и преобразовав ее каким-нибудь удобным способом в целое число, пользуясь при этом (возможно усложненной) поправкой на небольшой процент времени. Можно ли подобным образом реализовать биномиальное распределение для больших t?

23. [BM23] (Дж. фон Нейман.) Эквивалентны ли следующие два способа выработки случайной величины X (т. е. будет ли величина X иметь одно и то же распределение)?

Метод 1. Установить $X \leftarrow \sin((\pi/2)U)$, где U—случайное число.

Метод 2. Выработать два независимых случайных числа U, V и, если $U^2 + V^2 \ge 1$, повторить то же самое, пока не будет выполняться $U^2 + V^2 < 1$. Тогда положить $X \leftarrow |U^2 - V^2|/(U^2 + V^2)$.

24. [BM40] (С. Улам, Дж. фон Нейман.) Пусть V_0—случайное число между 0 и 1. Определим последовательность $\langle V_n\rangle$ соотношением $V_{n+1} = 4V_n \times (1 - V_n)$. Теперь, если вычисления делаются абсолютно точно, результат имеет распределение $\sin^2 \pi U$, где U—равномерно распределенное случайное число. Другими словами, функция распределения такова:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \frac{dx}{\sqrt{x(1-x)}}.$$

В самом деле, если мы напишем $V_n = \sin^2 \pi U_n$, то ясно, что $U_{n+1} = (2U_n) \bmod 1$. И из того, что почти все действительные числа имеют случайное двоичное представление (см. § 3.5), следует, что последовательность U_n равномерно распределенная. Но если вычисление V_n производится с конечной точностью, эти аргументы, оказываются неверными, так как скоро мы начинаем иметь дело с шумом от ошибок округления (von Neumann, *Collected Works,* Vol. V, pp. 768–770). Проведите теоретическое и экспериментальное (при разных значениях V_0) исследование последовательности $\langle V_n\rangle$, определенной выше, когда вычисления проводятся с конечной точностью. Похоже ли распределение на ожидаемое? Можно ли как-нибудь использовать эти числа?

25. [M25] Пусть X_1, X_2 …, X_5—двоичные слова, а каждый из двоичных разрядов независимо принимает значение 0 или 1 с вероятностью $1/2$. Какова вероятность того, что в данной позиции результат $X_1 \vee (X_2 \wedge (X_3 \vee (X_4 \wedge X_5)))$ содержит 1. Сделайте обобщение.

3.4.2. Случайная выборка и перемешивание[12]

При обработке данных часто бывает необходимо случайным образом и беспристрастно выбрать n записей из файла, в котором содержится N записей. Такая задача возникает, например, при контроле качества или в других статистических вычислениях, где требуются выборки. Обычно N очень велико, так что одновременно хранить все данные в памяти невозможно, поэтому нужно найти такую эффективную процедуру выбора n записей, которая позволила бы сразу решать, принять или отклонить каждую проходящую запись.

Для решения этой задачи было предложено несколько методов. Наиболее очевиден такой подход, когда любая запись выбирается с одной и той же вероятностью n/N. Иногда этот способ оказывается удобным, но при его использовании в выборке получается n записей только в *среднем,* причем стандартное отклонение равно $\sqrt{n(1-(n/N))}$: выборка может оказаться или слишком большой, или слишком малой для достижения желаемых результатов.

---

[12] В оригинале shuffling—тасование.—*Прим. перев.*

Приведем простую модификацию этого "очевидного" метода, лишенную такого недостатка. Если m записей уже было отобрано, мы должны включить $(t+1)$-ю запись в выборку с вероятностью $(n-m)/(N-t)$. Эта вероятность выражается именно такой величиной, поскольку из всех возможных способов выбора n записей из N таким образом, что m из них отбираются из первы t, в точности

$$\binom{N-t-1}{n-m-1} \Big/ \binom{N-t}{n-m} = \frac{n-m}{N-t} \tag{1}$$

доля способов выбирает $(t+1)$-й элемент.

Высказанную идею можно оформить в виде следующего алгоритма:

Алгоритм S. ***(Метод выборки.)*** Выбрать случайным образом n записей из N, где $0 < n \le N$.

S1 **[Начальная установка.]** Установить $t \leftarrow 0$, $m \leftarrow 0$.

S2 **[Выработать U.]** Выработать случайное число U, равномерно распределенное между нулем и единицей.

S3 **[Проверить.]** Если $(N-t)U \ge n-m$, перейти к **S5**.

S4 **[Отобрать.]** Включить запись в выборку и увеличить m и t на 1. Если $m < n$, перейти к **S2**, в противном случае выборка сделана и алгоритм закончен.

S5 **[Пропустить.]** Пропустить следующую запись (не включать ее в выборку), увеличить t на 1 и перейти к **S2**. ∎

На первый взгляд этот алгоритм может показаться ненадежным и даже неправильным, но внимательный анализ (см. упражнения, помещенные ниже) показывает, что он абсолютно верен. Нетрудно убедиться, что

(a) отбирается *ровно* n элементов,

(b) выборка совершенно случайна: в частности, вероятность выбора любого заданного элемента, например последнего элемента файла, равна n/N.

Утверждение (b) справедливо несмотря на то, что мы *не* выбираем $(t+1)$-й элемент с вероятностью n/N—мы выбираем его в соответствии с формулой (1)! Это привело к некоторой путанице в публикациях. Может ли читатель объяснить это кажущееся противоречие?

(*Замечание.* Для того чтобы избежать корреляций между выборками, полученными при различных просчетах по алгоритму S, следует брать различные источники случайных чисел U. Для этого, например, можно брать различные X_0 в линейном конгруэнтном методе. Значение X_0 может быть связано с календарной датой, или с последним значением X, полученным в предыдущем просчете.)

Обычно нам не нужно будет просматривать все N записей; действительно, поскольку в (b) указано, что последняя запись выбирается с вероятностью n/N, в $(1-n/N)$ доле всех случаев алгоритм закончит свою работу *прежде,* чем дойдет до последней записи. Среднее число рассмотренных записей для $n=2$ составляет около $(2/3)N$; общие формулы даны в упр. 5 и 6.

Алгоритм S применим лишь в том случае, когда величина N заранее известна. Предположим теперь, что мы хотим случайным образом выбрать n элементов из файла, но не знаем, из скольких элементов он состоит. Можно сначала пересчитать все элементы, а потом, чтобы сделать выборку, перебрать их вторично. Однако, вообще говоря, лучше сразу при пересчете отобрать $m \ge n$ элементов, где m гораздо меньше N, с тем, чтобы второй раз перебирать только m элементов. Весь фокус, разумеется, заключается в том, чтобы в результате получить истинно случайную выборку из первоначального файла.

Сейчас мы опишем остроумный прием, реализующий эту довольно неправдоподобную идею. Допустим, имеется датчик случайных чисел, способный породить по крайней мере N различных значений, прежде чем какое-либо значение повторится, например—это линейная конгруэнтная последовательность с периодом, большим N. (Длины периодов обычно составляют несколько миллиардов.) Для датчика можно рекомендовать алгоритм 3.2.2M, предложенный Маклареном и Марсальей.

Идея состоит в том, чтобы вычислить N случайных величин и установить, какие n из них являются наибольшими. Соответствующие им n записей и составят окончательную выборку. Во время первого перебора мы создаем "резервуар", в который входят только такие m записей—возможные кандидаты в выборку, что перед соответствующими им случайными величинами нет n больших значений (первые n элементов всегда попадают в резервуар).

Алгоритм R. ***(Выборка с резервуаром.)*** Для того чтобы формализовать описанный выше процесс, введем вспомогательную таблицу из n элементов вида (Y, I), где Y—случайная величина, соответствующая I-й записи в резервуаре. Вначале эта таблица целиком заполнена элементами $(0, 0)$.

R1 **[Начальная установка.]** Установить $m \leftarrow 0$.

R2 **[Выработать U.]** Пусть U—случайная величина, равномерно распределенная между нулем и единицей. (Как объяснялось ранее, предполагается, что U *не равно* никакому из ранее полученных на этом шаге случайных чисел.)

R3 **[Проверить.]** Пусть наименьшее значение Y во вспомогательной таблице (Y, I) равно Y_0. Если $U < Y_0$, перейти к **R6**.

R4 **[Добавить в резервуар.]** Перенести следующую запись из файла в резервуар и увеличить m на 1.

R5 **[Обновить таблицу.]** Заменить элемент таблицы (Y_0, I_0) на (U, m). Перейти к **R7**.

R6 **[Пропустить.]** Пропустить следующую запись файла.

R7 **[Конец файла?]** Если в файле больше нет записей, перейти к **R8**, в противном случае—к **R2**.

R8 **[Второй перебор.]** Отсортировать элементы таблицы (Y, I) по I. Записи в резервуаре, соответствующие результирующим значениям $I_1 < I_2 < \ldots < I_n$, составляют окончательную выборку. ∎

Читатель должен проработать пример применения этого алгоритма в упр. 9.

Таблицу (Y, I) в алгоритме R можно составить таким образом, что поиск Y_0 на шаге R3 будет очень быстрым и перенос (U, m) на шаге R5 также будет эффективным. За подробностями читатель отсылается к алгоритмам выборки деревьев в гл. 5.

При применении алгоритма R возникает естественный вопрос: Каков ожидаемый размер m резервуара?" Из упр. 11 следует, что среднее значение m равно $n(1 + H_N - H_n)$, что приближенно равно $n(1 + \ln(N/n))$. Следовательно, при $N/n = 1000$ в резервуаре будет приблизительно в 125 раз меньше элементов, чем в файле.

Заметим, что алгоритмы S и R можно использовать для того, чтобы получать выборки по нескольким независимым категориям одновременно. Например, имея большой файл имен и адресов жителей США, можно получить случайные выборки ровно по 10 человек каждого из 50 штатов, не перебирая 50 раз весь файл и не сортируя файл по штатам.

Алгоритмы S и R и несколько других методов получения выборок обсуждаются в статье Ч. Фаня, М. Маллера и И. Резухи [*Journal of the American Statistical Association,* **57** (1962), 387–402]. Отметим, что алгоритм S независимо разработал Т. Джоунс [*CACM,* **5** (1962), 343].

Задачу о выборке, как мы ее здесь описываем, можно рассматривать как вычисление случайного сочетания в соответствии с обычным определением сочетания из N элементов по n (см. п. 1.2.6). Рассмотрим теперь задачу вычисления случайной *перестановки* t объектов. Мы назовем ее *задачей перемешивания* (или тасования), поскольку тасование колоды из 52 карт есть случайная их перестановка.

Существуют два основных способа перемешивания. В первом, предложенном С. Уламом в начале пятидесятых годов используется, например, пять подпрограмм, каждая из которых производит определенную перестановку элементов $X_1, X_2, \ldots, X_t$. Предположим, что нужно получить *последовательность* случайных перестановок. Для получения очередной перестановки нашей последовательности мы действуем на предыдущую тремя из пяти перестановок. Эти три перестановки выбираются случайным образом и не обязательно отличаются друг от друга.

Этот метод похож на способ, которым люди обычно тасуют колоду карт (см. упр. 13). Может быть, читателю будет интересно узнать, какие перестановки получаются в колоде, перетасованной шулером. Для того чтобы это выяснить, автор попросил У. Логэна (специалиста по вычислительной технике и тасованию карт) зафиксировать результаты трех перемешиваний колоды из 52 карт. Получились следующие перестановки:

$\pi_1 = (1\,23\,31)\,(2\,42\,20)\,(3\,34\,52\,48\,24\,28\,39\,50\,15\,40\,41\,9\,17\,38\,25\,43\,47\,14\,8\,13\,35\,11\,22\,10\,12\,36\,49\,46\,44\,4\,37)$
$(5\,19\,18\,45)\,(6)\,(7\,30)\,(16\,51\,27\,32)\,(21\,29\,33\,26)$

$\pi_2 = (1\,5\,24\,15\,36\,19)\,(2\,18\,27\,52\,37\,44\,25\,47\,50\,46\,20\,7\,17\,10)$
$(3\,28\,6\,39\,13\,23\,41\,43\,11\,38\,49\,12\,4\,34\,8\,40\,21\,14\,30\,16\,45\,51\,22\,9\,33\,32\,35\,31\,26)\,(29\,42)\,(48)$

$\pi_3 = (1\,42\,19\,40\,30)\,(2\,51\,10\,34\,44\,48)\,(3\,12\,33\,49\,8\,23\,20\,25\,14\,7\,43\,36)$
$(4\,11\,26\,16\,46\,32\,24\,28\,9\,31\,29\,52\,21\,5\,47\,13\,39\,38\,17\,15\,37\,45\,6\,41\,50\,27)\,(18\,22\,35)$

Эти перестановки представлены с помощью циклов (ср. п. 1.3.3). Для того чтобы наиболее эффективным способом произвести перестановку последовательности чисел, следует использовать ее циклическую структуру. Если, например, использовать MIX, первая из приведенных выше перестановок

может быть запрограммирована как

```
01 LDA   X+1
02 LDX   X+31
03 STX   X+1
04 LDX   X+23
05 STX   X+31
06 STA   X+23
07 LDA   X+2
08 LDX   X+20
```

и т. д.

Мы хотим быть уверены в том, что если использовать наш набор из пяти заданных перестановок в надлежащей последовательности, то можно получить все $t!$ возможных перестановок. Можно показать, что, если две перестановки следующего вида:

$$\begin{aligned} &\pi_1 = (1\,2) && \text{(любое произведение непересекающихся циклов, каждый из них нечетной длины),}\\ &\pi_2 = (1) && \text{(любой цикл длины } t-1\text{, в который не входит элемент } 1\text{),} \end{aligned} \tag{2}$$

включены в набор, можно получить любую перестановку из t элементов путем последовательного применения π_1 и π_2 в некотором порядке (см. упр. 12.) Дж. Диксон показал, что почти $3/4$ из всех пар перестановок обладают этим свойством (*Math. Z.,* **110** (1969), 199–205).

Второй метод перемешивания тесно связан с алгоритмом 3.3.2P. Читатель должен вернуться к этому алгоритму, который дает простое соответствие между каждой из $t!$ возможных перестановок и набором чисел $C[1]$, $C[2]$, $\ldots$, $C[t]$, где $0 \le C[j] < j$. Можно легко получить случайный набор чисел, после чего с помощью упомянутого соответствия получить случайную перестановку.

Алгоритм P. ***(Перемешивание.)*** Пусть X_1, X_2, $\ldots$, X_t—набор t чисел, который нужно перемешать.

P1 [Начальная установка.] Установить $j \leftarrow t$.

P2 [Выработать U.] Выработать случайное число U, равномерно распределенное между нулем и единицей.

P3 [Обмен.] Установить $k \leftarrow \lfloor jU \rfloor + 1$. (Теперь k есть случайное целое число, расположенное в промежутке между 1 и j.) Взаимозаменить $X_k \leftrightarrow X_j$.

P4 [Уменьшить j.] Уменьшить j на 1. Если $j > 1$, возвратиться к шагу **P2.** ∎

Этот алгоритм впервые опубликовали Л. Мозес и Р. Оукфорд (Tables of Random Permutations (Stanford University Press, 1963)) и Р. Дурстенфелд (*CACM,* **7** (1964), 420). Алгоритм лучше, чем метод Улама, потому что он, в самом деле, вырабатывает "действительно случайные" перестановки и использует меньшую память.

Упражнения

1. [M12] Объясните формулу (1).
2. [20] Докажите, что при использовании алгоритма S отбираются точно n записей при условии, что $0 < n \le N$.
>3. [22] В алгоритме S $(t+1)$-й элемент выбирается с вероятностью $(n-m)/(N-t)$, а *не* n/N. В тексте, однако, написано, что выбор проводится беспристрастно, так что каждый элемент должен выбираться с *равной* вероятностью! Каким образом могут быть одновременно справедливы оба эти утверждения?
4. [M23] Пусть $p(m,t)$ есть вероятность того, что из первых t элементов выбирается ровно m. Пользуясь непосредственно алгоритмом S, покажите, что

$$p(m,t) = \binom{t}{m}\binom{N-t}{n-m}\Big/\binom{N}{n} \qquad \text{для } 0 \le t \le N.$$

5. [M24] Чему равно среднее значение t в момент завершения работы алгоритма S? (Другими словами, сколько записей из общего числа N мы переберем, прежде чем выборка будет закончена?)
6. [M24] Чему равно стандартное отклонение величины, вычисленной в предыдущем упражнении?
>7. [M25] Докажите, что при использовании алгоритма S любой *заданный* набор из n записей из общего числа N получается с вероятностью $1/\binom{N}{n}$. Тем самым будет показано, что выборка производится совершенно беспристрастно.

>8. [18] Что произойдет при работе алгоритма R, если в файле содержится меньше n записей? Предложите небольшую модификацию алгоритма для обнаружения такого случая.

9. [22] Предположим, что мы используем алгоритм R для $n = 3$, а файл состоит из 20 записей, и на шаге R2 были получены следующие случайные значения U:

$$0.53,\quad 0.97,\quad 0.66,\quad 0.30,\quad 0.81,\quad 0.19,\quad 0.09,\quad 0.31,\quad 0.67,\quad 0.62,$$
$$0.04,\quad 0.05,\quad 0.73,\quad 0.54,\quad 0.42,\quad 0.99,\quad 0.40,\quad 0.78,\quad 0.69,\quad 0.80;,$$

20 записей занумерованы от 1 до 20. Какие записи попадут в резервуар? Какие записи попадут в окончательную выборку?

10. [30] Каким образом нужно скомпоновать таблицу (Y, I) в алгоритме R (об этом упоминалось в тексте), чтобы эффективно выполнялись поиск и перенос на шагах R3 и R5?

>11. [M25] Предположим, что p_m есть вероятность того, что при первом переборе по алгоритму R в резервуар попадут ровно m элементов. Определите производящую функцию $G(z) = \sum_m p_m z^m$ и вычислите среднее и стандартное отклонение (используйте материал п. 1.2.10).

12. [M23] Пусть π_1, π_2—перестановки, определяемые соотношениями (2). Докажите справедливость следующих утверждений, (a) Некоторая степень π_1 равна циклу $(1\,2)$. (b) Некоторое произведение π_1 и π_2 равно циклу $(1\,k)$ для любого k, $2 \le k \le t$. (c) Любой цикл $(j\,k)$, $j \ne k$, может быть получен как произведение π_1 и π_2. (d) Все перестановки из t элементов могут быть получены как произведение π_1 и π_2.

13. [M23] (С. Голомб.) Обычный способ тасования карт состоит в следующем: колода делится'на две по возможности равные части, которые вставляются одна в другую. (В правилах карточных игр, опубликованных Хойлом, читаем: "Для того чтобы хорошо перемешать карты таким способом, нужно перетасовать их около трех раз".) Рассмотрим колоду из $2n-1$ карт $X_1, X_2, \ldots, X_{2n-1}$. Операция "идеального тасования" s разделяет колоду на две части $X_1, X_2, \ldots, X_n$ и $X_{n+1}, \ldots, X_{2n-1}$ и идеально смешивает их так, что получается последовательность $X_1, X_{n+1}, X_2, X_{n+2}, \ldots, X_{2n-1}, X_n$. Операция "снятия" c^j переводит последовательность $X_1, X_2, \ldots, X_{2n-1}$ в последовательность $X_{j+1}, \ldots, X_{2n-1}, X_1, \ldots, X_j$. Покажите, что при комбинировании "идеальных тасований" и "снятий" можно получить не более чем $(2n-1)(2n-2)$ различных последовательностей карт при условии, что $n > 1$.

3.5. * ЧТО ТАКОЕ СЛУЧАЙНАЯ ПОСЛЕДОВАТЕЛЬНОСТЬ?

A. Вводные замечания. В этой главе уже говорилось о том, как получать последовательности

$$\langle U_n \rangle = U_0, U_1, U_2, \ldots \tag{1}$$

действительных чисел, заключенных между нулем и единицей, т. е. таких, что $0 \le U_n < 1$. Эти последовательности назывались "случайными", хотя по способу получения они были совершенно детерминированными. Для того чтобы оправдать это название, мы писали, что числа "ведут себя так, как если бы они были действительно случайными". Для практических целей (в настоящее время) такого заявления может быть и достаточно, однако оно обходит один очень важный философский и теоретический вопрос: как точно сформулировать, что именно мы подразумеваем под "случайным поведением"? Нужно предложить количественное определение случайного поведения. Не следует пользоваться понятиями, которых по-настоящему не понимаешь, тем более что о случайных числах можно высказать много на первый взгляд парадоксальных утверждений.

Математическая статистика и теория вероятностей тщательно избегают ответа на наш вопрос, поскольку эти науки воздерживаются от абсолютных утверждений. Вместо этого рассматривается вопрос о том, какую *вероятность* следует приписать высказываниям, связанным со случайными последовательностями независимых событий. Аксиомы теории вероятностей позволяют без труда вычислять абстрактные вероятности, однако при этом остается неясным, что же в действительности означает понятие вероятности, или как это понятие можно осмысленно приложить к явлениям окружающего мира. Р. фон Мизес в книге "Вероятность, статистика и истина" (Probability, Statistics, and Truth, Macmillan, 1957) подробно обсуждает это положение и высказывает такую точку зрения, что определение вероятности зависит от того как определить случайную последовательность.

Приведем два описания понятия случайной последовательности, недавно предложенные двумя авторами.

> *Д. Х. Лемер (1951 г.):* "Случайная последовательность есть некое расплывчатое понятие, воплощающее идею последовательности, в которой каждый член непредсказуем для непосвященного и элементы которой удовлетворяют ряду традиционных среди статистиков

критериев, в известной степени зависящих от того, для каких применений служит эта последовательность".

Дж. Н. Фрэнклин (1962 г.): "Последовательность (1) случайна, если она обладает любым свойством, которым обладают все бесконечные последовательности независимых выборок случайных переменных из равномерного распределения"

Определение Фрэнклина существенно обобщает определение Лемера, поскольку оно требует, чтобы последовательность удовлетворяла *всем* статистическим критериям. Его определение не является абсолютно точным, и скоро мы убедимся в том, что разумная его интерпретация приводит к отрицанию существования такого объекта, как случайная последовательность! Таким образом, оно слишком ограничительно, поэтому попытаемся уточнить *определение Лемера.* Мы хотим получить относительно короткий перечень математических свойств, каждое из которых не противоречит нашему интуитивному представлению о случайной последовательности. Кроме того, этот перечень должен быть достаточно полным для того, чтобы *любую* последовательность, обладающую перечисленными свойствами, можно было бы отнести к "случайным". То, что мы разработаем в настоящем разделе, будет, по-видимому, удовлетворительным, с точки зрения приведенных выше соображений, определением случайности, хотя при этом останутся без ответа многие интересные вопросы.

Пусть u и v—действительные числа, $0 \le u < v \le 1$. Если U—случайная величина, равномерно распределенная между 0 и 1, то вероятность того, что $u \le U < v$, равна $v - u$. Например, если $u = 1/3$ и $v = 2/3$, вероятность того, что $1/3 \le U < 2/3$, равна $1/3$. Как обобщить это свойство на случай бесконечной последовательности U_0, U_1, U_2, ...? Очевидный ответ состоит в том, что если сосчитать, сколько раз U_n попадает в интервал между u и v, то среднее число попаданий должно быть равно величине $v - u$. Аналогичным образом вводилось интуитивное понятие вероятности: оно основывалось на частоте появления события. Если быть более точным, обозначим через $\nu(n)$ число значений j, $0 \le j < n$, таких, что $u \le U_j < v$. Мы хотим, чтобы отношение $\nu(n)/n$ стремилось к $v - u$ при стремлении n к бесконечности:

$$\lim_{n\to\infty} \nu(n)/n = v - u. \tag{2}$$

Если это условие удовлетворяется при любом выборе u и v, последовательность будем называть *равномерно распределенной.*

Обозначим через $S(n)$ некоторое утверждение относительно целого числа n и последовательности $U_1, U_2, \ldots$. Например, $S(n)$ может быть высказанным выше утверждением: "$u \le U_n < v$". Можно обобщить рассуждения, приведенные в предыдущем абзаце, и определить "вероятность того, что $S(n)$ истинно" по отношению к определенной бесконечной последовательности. Пусть $\nu(n)$—число значений j, $0 \le j < n$, таких, что $S(j)$ истинно.

Определение А. *Говорят, что* $\Pr(S(n)) = \lambda$, *если* $\lim_{n\to\infty} \nu(n)/n = \lambda$. ***(В словесной форме "вероятность того, что*** $S(n)$ ***истинно, равна*** λ, ***если предел*** $\nu(n)/n$ ***при*** n, ***стремящемся к бесконечности, есть*** λ***").***

Пользуясь этим определением, можно сказать, что последовательность U_0, U_1, ... равномерно распределена в том и только том случае, если $\Pr(u \le U_n < v) = v - u$ для всех действительных u, v, таких, что $0 \le u < v \le 1$.

Последовательность может быть равномерно распределенной, но не случайной. Если, например, U_0, U_1, ... и V_0, V_1, ...—две равномерно распределенные последовательности, то нетрудно показать, что последовательность

$$W_0, W_1, W_2, W_3, \ldots = \frac{1}{2}U_0, \frac{1}{2} + \frac{1}{2}V_0, \frac{1}{2}U_1, \frac{1}{2} + \frac{1}{2}V_1, \ldots \tag{3}$$

также равномерно распределена, поскольку последовательность $\frac{1}{2}U_0$, $\frac{1}{2}U_1$, ... равномерно распределена между 0 и $1/2$, а чередующаяся с ней последовательность $\frac{1}{2} + \frac{1}{2}V_0$, $\frac{1}{2} + \frac{1}{2}V_1$, ... равномерно распределена между $1/2$ и 1. В последовательности W за величиной, меньшей $1/2$, всегда следует величина, большая $1/2$, и обратно, так что эта последовательность не является случайной ни в каком разумном смысле. Нам нужно свойство более сильное, чем равномерная распределенность.

Пример, приведенный в предыдущем абзаце, приводит к естественному обобщению свойства равномерной распределенности, а именно к рассмотрению пар соседних элементов последовательности. Можно потребовать, чтобы для любых четырех чисел u_1, v_1, u_2, v_2, таких, что $0 \le u_1 < v_1 \le 1$, $0 \le u_2 < v_2 \le 1$, последовательность удовлетворяла требованию

$$\Pr(u_1 \le U_n < v_1 \text{ е } u_2 \le U_{n+1} < v_2) = (v_1 - u_1)(v_2 - u_2). \tag{4}$$

Более того, для любого положительного целого числа k можно ввести понятие k-*распределенной* последовательности.

Определение B. *Говорят, что последовательность* **(1)** k*-распределена, если*

$$\Pr(u_1 \le U_n < v_1, \ldots, u_k \le U_{n+k-1} < v_k) = (v_1 - u_1) \ldots (v_k - u_k) \tag{5}$$

для любых действительных чисел u_j, v_j, *таких, что* $0 \le u_j < v_j \le 1$ *при* $1 \le j \le k$.

Равномерно распределенная последовательность является 1-распределенной. Заметим, что если $k > 1$, то k-распределенная последовательность всегда $(k-1)$-распределена, поскольку в соотношении **(5)** можно положить $u_k = 0$ и $v_k = 1$. В частности, любая 4-распределенная последовательность является также и 3-распределенной, 2-распределенной и равномерно распределенной. Для заданной последовательности можно пытаться найти наибольшее k, такое, что последовательность k-распределена. Это приводит нас к следующему определению.

Определение C. *Последовательность называется* ∞*-распределенной, если она* k*-распределена, каково бы ни было положительное целое* k.

До сих пор мы рассматривали "последовательности на полуинтервале $[0, 1)$", т. е. последовательности действительных чисел, расположенных между нулем и единицей. Те же рассуждения приложимы и к последовательности целых чисел. Будем говорить, что последовательность $\langle X_n \rangle = X_0$, X_1, X_2, ... есть "b-ичная последовательность", если каждое из X_n есть одно из целых чисел 0, 1, ..., $b-1$. Таким образом, 2-ичная (двоичная) последовательность представляет собой последовательность нулей и единиц.

Заметим, что "b-ичное число" $x_1 x_2 \ldots x_k$ есть некоторый упорядоченный набор k целых чисел, причем $0 \le x_j < b$, где $1 \le j \le k$.

Определение D. *Назовем* b*-ичную последовательность* k*-распределенной, если*

$$\Pr(X_n X_{n+1} \ldots X_{n+k-1} = x_1 x_2 \ldots x_k) = 1/b^k \tag{6}$$

для всех b*-ичных чисел* $x_1 x_2 \ldots x_k$.

Из этого определения ясно, что если $U_0, U_1, \ldots$ есть k-распределенная последовательность на $[0, 1)$, то $\lfloor bU_0 \rfloor$, $\lfloor bU_1 \rfloor$, ... является k-распределенной b-ичной последовательностью. **(**В самом деле, если положить $u_j = x_j/b$, $v_j = (x_j + 1)/b$, $X_n = \lfloor bU_n \rfloor$, то формула **(5)** превратится в **(6).)** Более того, если b-ичная последовательность k-распределена, она также $(k-1)$-распределена: если сложить вероятности b-ичных чисел $x_1 \ldots x_{k-1}0$, $x_1 \ldots x_{k-1}1$, ..., $x_1 \ldots x_{k-1}(b-1)$, получится

$$\Pr(X_n \ldots X_{n+k-2} = x_1 \ldots x_{k-1}) = 1/b^{k-1}.$$

(Вероятности непересекающихся событий аддитивны, см. упр. 5.**)** Поэтому можно говорить о ∞-распределенной b-ичной последовательности, определив ее аналогично определению C.

Представление действительного положительного числа в b-ичной системе счисления можно рассматривать как b-ичную последовательность. Так, например, число π соответствует десятичной последовательности $3, 1, 4, 1, 5, 9, 2, 6, 5, 3, 5, 8, 9, \ldots$. Предполагают, что эта последовательность ∞-распределена, но никто пока не смог доказать, что она хотя бы 1-распределена.

Попробуем проанализировать введенные понятия более подробно в случае, когда k равно миллиону. В 1000000-распределенной двоичной последовательности будут попадаться отрезки, состоящие из миллиона нулей! Аналогично этому, в 1000000-распределенной на $[0, 1)$ последовательности будут попадаться отрезки длиной в миллион, состоящие из чисел, каждое из которых меньше половины. Правда, такие отрезки будут попадаться в среднем только в $(1/2)^{1000000}$ доле случаев, но важно то, что они *существуют*. Разумеется, то же самое может быть и в любой истинно случайной последовательности, если иметь в виду наше интуитивное понятие "истинно случайного". Легко себе представить, какую реакцию вызовет такой набор из миллиона "истинно случайных" чисел, использованный в вычислительном эксперименте; возникнут веские основания для жалобы на датчик случайных чисел! С другой стороны, если в последовательности чисел никогда не попадаются серии из миллиона U, каждое из которых меньше $1/2$, она не случайна и не будет годиться для других теоретически возможных приложений, в которых входными данными служат чрезвычайно длинные серии U. Подытоживая, можно сказать, что в *истинно случайной последовательности должна присутствовать локальная неслучайность*. Локальная неслучайность необходима в одних приложениях, но недопустима в других. Мы вынуждены заключить, что *ни одна последовательность "случайных" чисел не может отвечать требованиям, предъявляемым всеми приложениями*.

Точно так же имеются основания утверждать, что мы не можем судить о том, случайна ли *конечная* последовательность; каждая заданная последовательность ничем не хуже любой другой. Эти

соображения являются камнями преткновения на пути построения полезного определения случайности, но беспокоиться по этому поводу все-таки не следует. Можно дать такое определение случайности для бесконечных последовательностей действительных чисел, что соответствующая теория (надлежащим образом интерпретированная) будет весьма эффективна при рассмотрении тех обычных конечных последовательностей рациональных чисел, которые получаются на вычислительной машине. Более того, в этом разделе будет показано, что существует несколько внушающих доверие способов определения случайности конечных последовательностей.

B. ∞-распределенные последовательности. Изложим в сжатом виде теорию ∞-распределенных последовательностей. Нам придется пользоваться некоторыми результатами высшей математики, так что далее предполагается знакомство читателя с материалом курса математического анализа.

Во-первых, обобщим определение A, поскольку предел, фигурирующий в этом определении, существует не для всех последовательностей. Введем определения

$$\overline{\Pr}(S(n)) = \lim_{n\to\infty} \sup(\nu(n)/n), \quad \underline{\Pr}(S(n)) = \lim_{n\to\infty} \inf(\nu(n)/n). \tag{7}$$

Теперь величина $\Pr(S(n))$, если она имеет смысл, является общим значением величин $\underline{\Pr}(S(n))$ и $\overline{\Pr}(S(n))$.

Мы видели, что из k-распределенной на $[0,1)$ последовательности можно получить k-распределенную b-ичную последовательность, если U заменить на $\lfloor bU \rfloor$. Наша первая теорема показывает, что обратное утверждение также справедливо.

Теорема A. *Пусть $\langle U_n\rangle = U_0, U_1, U_2, \ldots$ —последовательность на $[0,1)$. Если*

$$\langle \lfloor b_j U_n \rfloor \rangle = \lfloor b_j U_0 \rfloor, \lfloor b_j U_1 \rfloor, \lfloor b_j U_2 \rfloor, \ldots$$

является k-распределенной b_j-ичной последовательностью для любого целого b_j, принадлежащего бесконечной последовательности $1 < b_1 < b_2 < b_3 < \ldots$, то $\langle U_n\rangle$ есть k-распределенная последовательность.

В качестве примера применения этой теоремы положим $b_j = 2^j$. Последовательность $\lfloor 2^j U_0 \rfloor$, $\lfloor 2^j U_1 \rfloor$, $\ldots$ есть не что иное, как последовательность первых j битов двоичного представления U_0, $U_1, \ldots$. Если все такие последовательности целых чисел k-распределены в смысле определения D, то последовательность действительных чисел $U_0, U_1, \ldots$ должна быть k-распределена в смысле определения B.

Доказательство теоремы A. Если последовательность $\lfloor bU_0 \rfloor, \lfloor bU_1 \rfloor, \ldots$ k-распределена, из аддитивности вероятностей следует, что соотношение (5) справедливо при условии, что u_j и v_j являются рациональными числами со знаменателем b. Пусть теперь u_j, v_j—любые действительные числа, а u'_j, v'_j—рациональные числа со знаменателем b, такие, что

$$u'_j \le u_j < u'_j + 1/b, \quad v'_j \le v_j < v'_j + 1/b.$$

Через $S(n)$ обозначим следующее утверждение:

$$u_1 \le U_n < v_1, \ldots, u_k \le U_{n+k-1} < v_k.$$

Мы имеем

$$\begin{aligned}
\overline{\Pr}(S(n)) &\le \Pr\left(u'_1 \le U_n < v'_1 + \frac{1}{b}, \ \ldots, u'_k \le U_{n+k-1} < v'_k + \frac{1}{b}\right) = \\
&= \left(v'_1 - u'_1 + \frac{1}{b}\right) \ldots \left(v'_k - u'_k + \frac{1}{b}\right); \\
\underline{\Pr}(S(n)) &\ge \Pr\left(u'_1 + \frac{1}{b} \le U_n < v'_1, \ldots, u'_k + \frac{1}{b} \le U_{n+k-1} < v'_k\right) = \\
&= \left(v'_1 - u'_1 - \frac{1}{b}\right) \ldots \left(v'_k - u'_k - \frac{1}{b}\right).
\end{aligned}$$

Заметим, что $\left|(v'_j - u'_j \pm 1/b) - (v_j - u_j)\right| \le 2/b$. Неравенства справедливы для всех $b = b_j$. При $j \to \infty$ имеем $b_j \to \infty$ и, таким образом,

$$\begin{aligned}
(v_1 - u_1) \ldots (v_k - u_k) &\le \underline{\Pr}(S(n)) \le \\
&\le \overline{\Pr}(S(n)) \le (v_1 - u_1) \ldots (v_k - u_k).
\end{aligned}$$

∎

Следующая теорема послужит основным орудием исследования k-распределенных последовательностей.

Теорема B. *Пусть $\langle U_n\rangle$—k – распределенная на $[0,1)$ последовательность, и $f(x_1,x_2,\ \ldots,x_k)$— интегрируемая в смысле Римана функция k переменных; тогда*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{0\le j<n} f(U_j,U_{j+1},\ \ldots,U_{j+k-1}) = \int_0^1\ldots\int_0^1 f(x_1,x_2,\ \ldots,x_k)\,dx_1\ldots dx_k. \tag{8}$$

Доказательство. Из определения k-распределенной последовательности следует, что этот результат справедлив в частном случае, когда

$$f(x_1,\ldots,x_k)=\begin{cases}1, & \text{если } u_1\le x_1<v_1,\ \ldots,\ u_k\le x_k<v_k,\\ 0 & \text{в противном случае.}\end{cases} \tag{9}$$

Значит, соотношение (8) справедливо, если $f=a_1f_1+a_2f_2+\cdots+a_mf_m$, где f_j являются функциями вида (9). Другими словами, соотношение (8) справедливо, если f—"ступенчатая функция", постоянная внутри каждой части единичного k-мерного куба, полученной разбиением этого куба плоскостями, параллельными координатным осям.

Пусть теперь f—любая интегрируемая в смысле Римана функция. Мы знаем (из определения интегрируемости в смысле Римана), что если ε—любое положительное число, то существуют ступенчатые функции $\underline{f}$ и $\overline{f}$, такие, что $\underline{f}(x_1,\ \ldots,x_k)\le f(x_1,\ \ldots,x_k)\le\overline{f}(x_1,\ \ldots,x_k)$, и разности между интегралами от $\underline{f}$, f и $\overline{f}$ будут меньше ε. Поскольку (8) справедливо для $\underline{f}$ и $\overline{f}$ и

$$\begin{aligned}\frac{1}{n}\sum_{0\le j<n}\underline{f}(U_j,\ \ldots,U_{j+k-1}) &\le \frac{1}{n}\sum_{0\le j<n} f(U_j,\ \ldots,U_{j+k-1})\le\\ &\le\frac{1}{n}\sum_{0\le j<n}\overline{f}(U_j,\ \ldots,U_{j+k-1}),\end{aligned}$$

мы получаем, что (8) справедливо также и для f. ∎

В качестве первого приложения этой теоремы рассмотрим *проверку перестановок*, описанную в п. 3.3.2. Пусть $(p_1,p_2,\ \ldots,p_k)$ есть любая перестановка чисел $1, 2, \ldots, k$. Мы хотим показать, что

$$\Pr(U_{n+p_1-1}<U_{n+p_2-1}<\ldots<U_{n+p_k-1})=\frac{1}{k!}. \tag{10}$$

Для этого предположим, что последовательность $\langle U_k\rangle$ k-распределена, и положим

$$f(x_1,\ldots,x_k)=\begin{cases}1, & \text{если } x_{p_1}\le x_{p_2}\le\cdots\le x_{p_k},\\ 0 & \text{в противном случае.}\end{cases}$$

Имеем

$$\begin{aligned}\Pr(U_{n+p_1-1}<U_{n+p_2-1}<\cdots<U_{n+p_k-1}) &=\\ &=\int_0^1\ldots\int_0^1 f(x_1,\ldots,x_k)\,dx_1\ldots dx_k\\ &=\int_0^1 dx_{p_k}\int_0^{x_{p_k}}\ldots\int_0^{x_{p_3}}dx_{p_2}\int_0^{x_{p_2}}dx_{p_1}=\frac{1}{k!}\end{aligned}$$

Следствие P. *Если последовательность k-распределена на $[0,1)$, то она удовлетворяет проверке перестановок порядка k в смысле соотношения (10).* ∎

Можно также показать, что она удовлетворяет *тесту последовательной корреляции.*

Следствие S. *Если последовательность $(k+1)$-распределена на $[0,1)$, то коэффициент последовательной корреляции между U_n и U_{n+k} стремится к нулю:*

$$\lim_{n\to\infty}\frac{\frac{1}{n}\sum U_jU_{j+k}-\left(\frac{1}{n}\sum U_j\right)\left(\frac{1}{n}\sum U_{j+k}\right)}{\sqrt{\left(\frac{1}{n}\sum U_j^2-\left(\frac{1}{n}\sum U_j\right)^2\right)\left(\frac{1}{n}\sum U_{j+k}^2-\left(\frac{1}{n}\sum U_{j+k}\right)^2\right)}}=0.$$

(Для всех суммирований здесь $0 \le j < n$*.)*

Доказательство. Из теоремы B следует, что величины

$$\frac{1}{n}\sum U_j U_{j+k}, \quad \frac{1}{n}\sum U_j^2, \quad \frac{1}{n}\sum U_{j+k}^2, \quad \frac{1}{n}\sum U_j, \quad \frac{1}{n}\sum U_{j+k}$$

при $n \to \infty$ стремятся соответственно к пределам $1/4$, $1/3$, $1/3$, $1/2$, $1/2$. ∎

Рассмотрим теперь некоторые более общие свойства последовательностей. Мы определили свойство k-распределенности для всех групп из k стоящих рядом элементов последовательности. Например, последовательность является 2-распределенной тогда и только тогда, когда точки

$$(U_0, U_1), \quad (U_1, U_2), \quad (U_2, U_3), \quad (U_3, U_4), \quad (U_4, U_5), \quad \ldots$$

равномерно распределены в единичном квадрате. Однако вполне возможно, что пары точек, взятые через одну, не будут равномерно распределены. При этом недостача точек (U_{2n-1}, U_{2n}) в некоторой области может компенсироваться точками (U_{2n}, U_{2n+1}). Ясно, например, что периодическая двоичная последовательность

$$\langle X_n \rangle = 0,0,0,1,\ 0,0,0,1,\ 1,1,0,1,\ 1,1,0,1,\ 0,0,0,1,\ \ldots \tag{11}$$

с периодом, равным 16, 3-распределена, однако в подпоследовательности элементов с четными номерами $\langle X_{2n} \rangle = 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 1,\ 0,\ \ldots$ в три раза больше нулей, чем единиц, в то время как в подпоследовательности элементов с нечетными номерами $\langle X_{2n+1} \rangle = 0,\ 1,\ 0,\ 1,\ 1,\ 1,\ 1,\ 1,\ \ldots$ в три раза больше единиц, чем нулей.

Из приведенного выше примера следует, что если последовательность $\langle U_n \rangle$ ∞-распределена, то совсем не очевидно, что подпоследовательность $\langle U_{2n} \rangle = U_0, U_2, U_4, U_6, \ldots$ ∞-распределена или даже 1-распределена. Мы увидим, однако, что $\langle U_{2n} \rangle$ действительно ∞-распределена и что справедливо даже более сильное утверждение.

Определение E. *Говорят, что последовательность* $\langle U_n \rangle$ (m, k)*-распределена на* $[0, 1)$*, если*

$$\Pr(u_1 \le U_{mn+j} < v1, u_2 \le U_{mn+j+1} < v_2, \ldots, u_k \le U_{mn+j+k-1} < v_k) = (v_1 - u_1)\ldots(v_k - u_k)$$

при любом выборе действительных чисел u_r*,* v_r*, таких, что* $0 \le u_r < v_r \le 1$*, при* $1 \le r \le k$ *и для всех целых* j*, таких, что* $0 \le j < m$*.*

В частном случае $m = 1$ из определения E следует, что $\langle U_n \rangle$—k-распределенная последовательность; когда $m = 2$, это значит, что группы из k элементов, начинающиеся с элемента с четным номером, должны иметь такую же плотность, как и начинающиеся с нечетного номера, и т. д.

Некоторые свойства последовательностей, удовлетворяющих определению E, очевидны:

(m, k)-распределенная последовательность (m, κ)-распределена при $1 \le \kappa \le k$. (12)

(m, k)-распределенная последовательность (d, k)-распределена для всех делителей d числа m. (13)

Аналогично тому, как это сделано выше (определение D), можно определить понятие (m, k)-распределенной b-ичной последовательности. Доказательство теоремы A при этом остается в силе и для (m, k)-распределенных последовательностей.

Из следующей теоремы, во многих отношениях удивительной, вытекает, что свойство ∞-распределенности является гораздо более сильным, чем мы могли предполагать, вводя это определение.

Теорема C. *(A. Нивен и X. Цукерман.)* ∞*-распределенная последовательность является* (m, k)*-распределенной для любых положительных целых* m *и* k*.*

Доказательство. Достаточно доказать теорему для b-ичных последовательностей, с помощью только что упомянутого обобщения теоремы A. Более того, можно считать, что $m = k$, поскольку, вследствие утверждений (12) и (13), последовательность является (m, k)-распределенной, если она (mk, mk)-распределена.

Таким образом, мы докажем, что *любая* ∞*-распределенная* b*-ичная последовательность* X_0, $X_1, \ldots$ (m, m)*-распределена для всех целых положительных* m. Приведем упрощенный вариант доказательства, опубликованного Нивеном и Цукерманом (*Pacific Journal of Mathematics,* **1 (1951), 103–109).**

Доказательство теоремы основано на важной идее, используемой во многих математических рассуждениях: "Если значения суммы m величин и суммы их квадратов не противоречат гипотезе о том, что эти m величин равны, то эта гипотеза верна". Сильную форму этого принципа дает

Лемма E. *Пусть заданы m последовательностей чисел $\langle y_{jn}\rangle = y_{j0}, y_{j1}, y_{j2}, \ldots$, где $1 \le j \le m$. Предположим, что*

$$\begin{aligned} \lim_{n\to\infty}(y_{1n} + y_{2n} + \cdots + y_{mn}) &= m\alpha, \\ \lim_{n\to\infty}\sup(y_{1n}^2 + y_{2n}^2 + \cdots + y_{mn}^2) &\le m\alpha^2. \end{aligned} \tag{14}$$

Тогда для каждого j существует $\lim_{n\to\infty} y_{jn}$, и он равен α.

Необычайно простое доказательство этой леммы дано в упр. 9.∎

Теперь продолжим доказательство теоремы C. Пусть $x = x_1x_2\ldots x_m$ есть b-ичное число. Будем говорить, что x *появляется* на p-м месте последовательности, если $X_{p-m+1}X_{p-m+2}\ldots X_p = x$. Пусть $\nu_j(n)$ обозначает число появлений x на p-м месте при условии, что $p < n$ и $p \bmod m = j$. Пусть $y_{jn} = \nu_j(n)/n$. Мы хотим показать, что

$$\lim_{n\to\infty} y_{jn} = 1/mb^m. \tag{15}$$

Прежде всего заметим, что

$$\lim_{n\to\infty}(y_{0n} + y_{1n} + \cdots + y_{(m-1)n}) = 1/b^m, \tag{16}$$

поскольку последовательность m-распределена. Используя лемму E и соотношение (16), мы докажем теорему, если сумеем показать, что

$$\lim_{n\to\infty}\sup(y_{0n}^2 + y_{1n}^2 + \cdots + y_{(m-1)n}^2) \le 1/mb^{2m}. \tag{17}$$

Это неравенство не очевидно, и, чтобы доказать его справедливость, необходимо провести довольно тонкие преобразования. Пусть q кратно m; рассмотрим выражение

$$C(n) = \sum_{0\le j<m} \binom{\nu_j(n) - \nu_j(n-q)}{2}. \tag{18}$$

Это число пар появлений x на местах p_1, p_2 при условии $n - q \le p_1 < p_2 < n$ с $p_2 - p_1$, кратным m. Теперь рассмотрим сумму

$$S_N = \sum_{1\le n\le N+q} C(n) \tag{19}$$

При вычислении S_N каждая пара появлений x на местах p_1, p_2 при условии $p_1 < p_2 < p_1 + q$ с $p_2 - p_1$, кратным m, и с $p_1 \le N$ учитывается $p_1 + q - p_2$ раз (именно, когда $p_2 < n \le p_1 + q$), и пары таких появлений с $N < p_1 < p_2 < N + q$ считаются $N + q - p_2$ раз.

Пусть $d_t(n)$ есть число пар появлений x на местах p_1, p_2 при условии $p_1 + t = p_2 < n$. Из рассуждении, приведенных выше, имеем

$$\sum_{0<t<q/m} (q - mt)d_{mt}(N + q) \ge S_n \ge \sum_{0<t<q/m} (q - mt)d_{mt}(N). \tag{20}$$

Первоначальная последовательность q-распределена, следовательно,

$$\lim_{N\to\infty} \frac{1}{N} d_{mt}(N) = 1/b^{2m} \tag{21}$$

для всех $0 < t < q/m$, и поэтому из (20) вытекает

$$\lim_{N\to\infty} \frac{S_N}{N} = \sum_{0<t<q/m} (q - mt)/b^{2m} = q(q - m)/2mb^{2m}. \tag{22}$$

Отсюда, после некоторых преобразований, убедимся в справедливости теоремы.

По определению

$$2S_N = \sum_{1\le n\le N+q} \sum_{0\le j<m} ((\nu_j(n) - \nu_j(n-q))^2 - (\nu_j(n) - \nu_j(n-q))),$$

и если мы устраним члены, не возведенные в квадрат, то, используя (16), получим

$$\lim_{N\to\infty} \frac{T_N}{N} = q(q - m)/mb^{2m} + q/b^m, \tag{23}$$

где

$$T_N = \sum_{1\le n\le N+q} \sum_{0\le j<m} (\nu_j(n) - \nu_j(n-q))^2.$$

По неравенству

$$\frac{1}{r}\left(\sum_{1\le j\le r} a_j\right)^2 \le \sum_{1\le j\le r} a_j^2$$

(ср. с упр. 1.2.3-30) получаем, что

$$\lim_{N\to\infty} \sup \sum_{0\le j<m} \frac{1}{N(N+q)} \left(\sum_{1\le n\le N+q} (\nu_j(n) - \nu_j(n-q))\right)^2 \le q(q-m)/mb^{2m} + q/b^m. \tag{24}$$

Имеем также

$$q\nu_j(N) \le \sum_{N<n\le N+q} \nu_j(n) = \sum_{1\le n\le N+q} (\nu_j(n) - \nu_j(n-q)) \le q\nu_j(N+q)$$

и, подставляя это в (24), получаем

$$\lim_{N\to\infty} \sup \sum_{0\le j<m} (\nu_j(N)/N)^2 \le (q-m)/qmb^{2m} + 1/qb^m. \tag{25}$$

Мы доказали справедливость этой формулы для случая q, кратного m, и если теперь устремить $q\to\infty$, получим (17), тем самым завершив доказательство.

Возможно, что доказательство, приведенное в статье Дж. Касселса (*Pacific Journal of Mathematics,* **2** (1952), 555–557), проще. ∎

Нетривиальность этой теоремы видна из упр. 29 и 30. Из них же следует тот факт, что, как показано в (25), вероятности, связанные с q-распределенной последовательностью, отличаются от вероятностей, связанных с (m,m)-распределенной, не более чем на величину порядка $1/\sqrt{q}$. Для справедливости теоремы необходимо неослабленное предположение о ∞-распределенности.

Как следствие теоремы C можно доказать, что ∞-распределенная последовательность удовлетворяет проверке серий, тесту "наибольшее из t" и проверке подпоследовательностей, упомянутым в п. 3.3.2. Нетрудно показать, что удовлетворяются также проверка интервалов, покер-тест (проверка комбинаций) и проверка на монотонность (см. упр. с 12 по 14). Проверить тест собирателя купонов труднее, однако он также удовлетворяется (см. упр. 15 и 16).

Следующая теорема гарантирует существование относительно простых ∞-распределенных последовательностей.

Теорема F. *(Дж. Фрэнклин.) Последовательность* $U_0, U_1, \ldots$ *на* $[0,1)$, *где*

$$U_n = (\theta^n \bmod 1), \tag{26}$$

∞*-распределена для почти всех действительных чисел* $\theta > 1$. *Другими словами, множество*

$$\{\,\theta \mid \theta > 1 \text{ и последовательность (26) не } \infty\text{-распределена}\,\}$$

имеет меру нуль.

Доказательства этой теоремы и некоторых ее обобщений приведены в упоминаемой ниже статье Фрэнклина. ∎

Фрэнклин показал, что для того, чтобы последовательность (26) была ∞-распределена, θ должно быть трансцендентным числом. Хотя известно, что последовательность (26) ∞-распределена для *почти всех* чисел θ, мы не знаем *ни одного* конкретного θ, для которого это справедливо. С помощью трудоемких вычислений с многократно увеличенной точностью были получены степени $(\pi^n \bmod 1)$ при $n \le 10\,000$. Старшие 35 битов каждого из этих чисел были записаны на диск и успешно использовались как источник случайных чисел. Интересно, что хотя из теоремы F следует, что вероятность того, что последовательность степеней $(\pi^n \bmod 1)$ ∞-распределена, равна 1, однако, поскольку множество действительных чисел несчетно, мы не можем из этого заключить, что она ∞-распределена. Можно быть уверенным в том, что в течение нашей жизни никто не *докажет,* что эта последовательность *не* является ∞-распределенной, однако, возможно, так оно и есть на самом деле. В связи с изложенным

возникает вопрос о том, существует ли ∞-распределенная последовательность, которую можно выписать в *явном* виде; иными словами, *существует ли алгоритм, по которому для всех* $n \geq 0$ *можно вычислить действительные числа* U_n*, так что последовательность* $\langle U_n \rangle$ *будет* ∞*-распределенной?* Оказывается, такой алгоритм существует, что видно, например, из статьи автора "Construction of a Random Sequence" (*BIT*, **5** (1965), 246–250). Последовательность, построенная в этой статье, целиком состоит из рациональных чисел. Каждое число U_n имеет конечное представление в двоичной системе счисления. Несколько более сложный явный способ построения ∞-распределенной последовательности можно получить из приведенной ниже теоремы W.

C. Эквивалентны ли понятия ∞-распределенности и случайности? Из всего сказанного выше про ∞-распределенные последовательности следует, что понятие ∞-распределенности важно само по себе. Можно также с достаточным основанием считать, что следующее определение хорошо описывает интуитивное понятие случайности.

Определение R1. *Последовательность на* $[0, 1)$ *называется "случайной", если она* ∞*-распределена.*

Мы видели, что такие последовательности удовлетворяют всем тестам п. 3.3.2 и еще многим другим.

Попробуем критически подойти к этому определению. Прежде всего, является ли любая "истинно случайная" последовательность ∞-распределенной? Существует несчетное количество последовательностей U_0, U_1, ... действительных чисел, заключенных между нулем и единицей. Если значения U_0, U_1, ... получаются с помощью датчика истинно случайных чисел, любую из последовательностей можно считать равноценной. При этом некоторые из этих последовательностей (в действительности бесконечно большое их число) не будут даже равномерно распределены. С другой стороны, при любом разумном определении вероятности на этом пространстве всех возможных последовательностей мы должны заключить, что случайная последовательность ∞-распределена с *вероятностью* 1. Таким образом, мы приходим к формализации определения случайности, данного Фрэнклином (см. начало параграфа).

Определение R2. *Последовательность* $\langle U_n \rangle$ *на* $[0, 1)$ *называется "случайной", если для любого свойства* P*, такого, что* $P(\langle V_n \rangle)$ *справедливо с вероятностью 1 для последовательности* $\langle V_n \rangle$ *независимых выборок случайных величин из равномерного распределения, справедливо* $P(\langle U_n \rangle)$.

Возможно ли, что определение R1 эквивалентно определению R2? Попробуем выдвинуть против определения R1 некоторые возражения.

Прежде всего определение R1 имеет дело только с предельными свойствами последовательности при $n \to \infty$. Существуют ∞-распределенные последовательности, начинающиеся отрезком из миллиона нулей. Следует ли считать такую последовательность случайной?

Это возражение не очень серьезно. Пусть ε—любое положительное число, тогда вполне возможно, что каждый из первого миллиона членов последовательности меньше ε. Как мы уже отмечали ранее, нет способа, позволяющего судить о том, случайна или нет конечная последовательность. Истинно случайная последовательность с вероятностью единица содержит бесконечно много отрезков по миллиону последовательных членов, каждый из которых меньше ε. Почему же такой отрезок не может оказаться в начале последовательности?

С другой стороны, рассмотрим определение R2, и пусть P—свойство последовательности, состоящее в том, что все ее элементы различны. Свойство P справедливо с вероятностью единица, поэтому любая последовательность с миллионом нулей по *этому* критерию не является случайной.

Пусть теперь свойство P заключается в том, что *ни один* элемент последовательности не равен нулю. Оно справедливо с вероятностью единица, поэтому по определению R2 любая последовательность, у которой есть нулевой элемент, не является случайной. Рассмотрим более общий случай: пусть x_0—любое заданное число, заключенное между нулем и единицей, и P—свойство, состоящее в том, что ни один элемент последовательности не равен x_0. Из определения R2 следует, что никакая случайная последовательность не может содержать элемент, равный x_0! Теперь можно доказать, что *ни одна последовательность не может удовлетворять условиям, сформулированным в определении* R2. (В самом деле, если U_0, U_1, ... есть последовательность, удовлетворяющая этим условиям, то положим $x_0 = U_0$.)

Таким образом, если R1—слишком слабое определение, то R2—слишком сильное. "Правильное" определение должно быть менее ограничительным, чем R2. Однако мы, вообще говоря, еще не доказали, что R1 слишком слабо, поэтому продолжим его изучение. Выше говорилось о том, что можно построить ∞-распределенную последовательность *рациональных* чисел. (Разумеется, в этом нет ничего особенно удивительного: см. упр. 1.8.) Почти все действительные числа иррациональны, поэтому можно потребовать, чтобы для случайной последовательности имело место равенство

$$\Pr(U_n \text{ рационально}) = 0.$$

Заметим, что равномерная распределенность, по определению означает, что $\Pr(u \le U_n < v) = v - u$. Используя теорию меры, легко обобщить это определение: "Если мера множества $S \subseteq [0,1)$ равна μ, то

$$\Pr(U_n \in S) = \mu \tag{27}$$

для всех случайных последовательностей $\langle U_n \rangle$". В частности, если множество S составлено из рациональных чисел, оно имеет меру нуль, и, таким образом, никакая последовательность рациональных чисел не является равномерно распределенной в этом обобщенном смысле. Можно ожидать, что теорема B обобщается на случай интегрирования в смысле Лебега, если потребовать выполнения свойства (27). Однако мы опять приходим к выводу, что определение (27) является слишком жестким, поскольку ни одна последовательность *не* обладает этим свойством! Если $U_0, U_1, \ldots$ —некоторая последовательность, то мера множества $S = \{ U_0, U_1, \ldots \}$ равна нулю, однако $\Pr(U_n \in S) = 1$. Таким образом, путем тех же рассуждений, с помощью которых мы исключили из случайных последовательностей рациональные числа, можно исключить все случайные последовательности.

Пока все еще не показано, что определение R1 непригодно. Против него, однако, существуют веские возражения. Если, например, имеется случайная в интуитивном смысле последовательность, бесконечная ее подпоследовательность

$$U_0, U_1, U_4, U_9, \ldots, U_{n^2}, \ldots \tag{28}$$

должна также быть случайной. Для ∞-распределенной последовательности это не всегда справедливо. (Действительно, если взять любую ∞-распределенную последовательность и положить $U_{n^2} \leftarrow 0$ для всех n, величины $\nu_k(n)$, которые появляются при проверке k-распределенности, изменятся не больше, чем на величину порядка $\sqrt{n}$, так что пределы отношений $\nu_k(n)/n$ не изменятся.) Отсюда следует, что R1 не обладает этим свойством случайности.

Попробуем усилить R1 следующим образом.

Определение R3. *Последовательность на* $[0,1)$ *называется "случайной", если каждая ее бесконечная подпоследовательность* ∞*-распределена.*

Однако и это определение оказывается слишком ограничительным, поскольку из любой равномерно распределенной последовательности $\langle U_n \rangle$ можно выделить монотонную подпоследовательность $U_{s_0} < U_{s_1} < U_{s_2} < \ldots$.

Следует так ограничить класс рассматриваемых подпоследовательностей, чтобы каждый член последовательности, заранее неизвестный, можно было бы выбирать по некоторому правилу. Таким образом, получаем

Определение R4. *Последовательность* $\langle U_n \rangle$ *на* $[0,1)$ *называется "случайной", если каков бы ни был эффективный алгоритм, с помощью которого получается бесконечная последовательность различных неотрицательных целых чисел* s_n*, где* $n \ge 0$*, подпоследовательность* $U_{s_0}, U_{s_1}, \ldots$*, соответствующая этому алгоритму, является* ∞*-распределенной.*

Алгоритмы, о которых идет речь,—это процедуры, вычисляющие s_n по заданному n. В этом определении говорится обо всех бесконечных "рекурсивно перечислимых" подпоследовательностях в соответствии с обычным определением рекурсивной перечислимости (см. гл. 11).

Относительно данного определения следует сделать несколько замечаний. Последовательность $\langle \pi^n \bmod 1 \rangle$ наверняка *не* будет удовлетворять определению R4, поскольку или она не является равномерно распределенной, или существует эффективный алгоритм, определяющий бесконечную последовательность s_n, такую, что $(\pi^{s_0} \bmod 1) < (\pi^{s_1} \bmod 1) < \ldots$. Можно поэтому утверждать, что *ни одна явным образом определенная последовательность не может удовлетворять определению* R4. С этим следует согласиться, если считать, что явным образом определенная последовательность не может быть действительно случайной. Вполне возможно, однако, что последовательность $\langle \theta^n \bmod 1 \rangle$ будет удовлетворять определению R4 для почти всех действительных чисел $\theta > 1$. Здесь нет противоречия, поскольку почти все θ невозможно вычислить с помощью алгоритма. Известны, например, следующие факты. (i) Последовательность $\langle \theta^m \bmod 1 \rangle$ удовлетворяет определению R4 для почти всех действительных $\theta > 1$, если условие ∞-распределенности заменить на условие 1-распределенности. Эта теорема была доказана Ю. Коксмой (*Compositio Mathematica,* **2** (1935), 250–258). (ii) Последовательность $\langle \theta^{s(n)} \bmod 1 \rangle$ ∞-распределена для почти всех действительных $\theta > 1$, если $\langle s(n) \rangle$ есть последовательность целых чисел, таких, что $s(n+1) - s(n) \to \infty$ при $n \to \infty$. Можно, например, положить $s(n) = n^2$, или $s(n) = \lfloor n \log n \rfloor$.

Определение R4 намного сильнее определения R1, однако и оно все еще слишком слабо. Пусть, например, имеется истинно случайная последовательность $\langle U_n \rangle$. Определим подпоследовательность

$\langle U_{s_n}\rangle$ следующим образом: $s_0 = 0$, и для $n > 0$ число s_n—наименьшее положительное целое, такое, что все числа U_{s_n-1}, U_{s_n-2}, ..., U_{s_n-n} меньше половины. Тем самым мы рассматриваем подпоследовательность величин, следующих сразу же за серией из n членов, каждый из которых меньше $1/2$. Предположим, что "$U_n < 1/2$" соответствует выпадению "решетки" при бросании монеты. Игроки обычно считают, что если монета бросается честно, то длинная серия выпадения "решеток" увеличивает вероятность последующего выпадения "орла", и определенная нами подпоследовательность $\langle U_{s_n}\rangle$ соответствует стратегии игрока, при которой он делает n-ю ставку при бросании монеты, следующем вслед за первым последовательным выпадением n "решеток". Игрок может считать, что $\Pr(U_{s_n} \geq 1/2)$ превосходит половину, но, разумеется, в истинно случайной последовательности значения $\langle U_{s_n}\rangle$ будут совершенно случайными. Нет такой стратегии, которая обеспечивала бы преимущество в игре. В определении **R4** ничего не говорится о подпоследовательностях, составленных согласно такой стратегии, поэтому нужно предложить нечто большее.

Определим "правило построения подпоследовательности" $\mathcal{R}$ как бесконечную последовательность функций $\langle f_n(x_1, \ldots, x_n)\rangle$, где $n \geq 0$, f_n—функция n переменных, и $f_n(x_1, \ldots, x_n)$ может принимать значение 0 или 1. Здесь x_1, ..., x_n являются элементами некоторого множества S. **(**Таким образом, f_0 есть постоянная функция, равная 0 или 1.**)** Правило $\mathcal{R}$ построения подпоследовательности определяет подпоследовательность любой бесконечной последовательности $\langle X_n\rangle$ элементов S следующим образом: *n-й член X_n содержится в подпоследовательности $\langle X_n\rangle\,\mathcal{R}$ в том и только том случае, когда* $f_n(X_0, X_1, \ldots, X_{n-1}) = 1$. Заметим, что определенная таким образом подпоследовательность $\langle X_n\rangle\,\mathcal{R}$ не обязательно является бесконечной и может даже быть пустой.

"Подпоследовательность игрока", описанная выше, соответствует следующему правилу построения подпоследовательности: "$f_0 = 1$; $f_n(x_1, \ldots, x_n) = 1$ при $n > 0$ в том и только том случае, когда существует некоторое k, $0 < k \leq n$, такое, что $x_m < 1/2$, $x_{m-1} < 1/2$, ..., $x_{m-k+1} < 1/2$, ..., $x_{m-k+1} < 1/2$ при $m = n$, но не в случае, когда $k \leq m < n$".

Правило построения подпоследовательности $\mathcal{R}$ называется *"вычислимым"*, если существует эффективный алгоритм, позволяющий определить значение $f_n(x_1, \ldots, x_n)$ по входным данным $n, x_1, \ldots, x_n$. Стремясь определить понятие случайности, нужно ограничиваться вычислимыми правилами построения подпоследовательностей, в противном случае мы получим слишком сильное определение, подобное **R3**. Произвольные действительные числа не могут служить входными данными эффективных алгоритмов, потому что если, например, действительное число x в десятичной системе счисления определяется бесконечной последовательностью цифр, то не существует алгоритма, по которому можно было бы определить, справедливо ли неравенство $x < 1/3$, поскольку пришлось бы исследовать все знаки числа $0,333\ldots$. В связи с этим вычислимые правила построения подпоследовательности не приложимы ко всем последовательностям "на $[0, 1)$", и следующее определение будет удобно основывать на b-ичных последовательностях.

Определение R5. *b-ичная последовательность называется "случайной", если каждая ее бесконечная подпоследовательность, определенная с помощью вычислимого правила построения подпоследовательности, 1-распределена.*

Последовательность $\langle U_n\rangle$ на $[0, 1)$ называется "случайной", если b-ичная последовательность $\langle \lfloor bU_n \rfloor\rangle$ "случайна" для всех целых чисел $b \geq 2$.

Заметим, что в определении **R5** подпоследовательность "1-распределена", а не "∞-распределена". Интересно заметить, что общность при этом не теряется. В самом деле, для любого b-ичного числа $a_1 \ldots a_k$ можно так определить очевидным образом вычислимое правило $\mathcal{R}(a_1 \ldots a_k)$ построения подпоследовательности. Положим $f_n(x_1, \ldots, x_n) = 1$ в том и только том случае когда $n \geq k - 1$ и $x_{n-k+1} = a_1$, ..., $x_{n-1} = a_{k-1}$, $x_n = a_k$. Если $\langle X_n\rangle$ является k-распределенной b-ичной последовательностью, то упомянутое правило $\mathcal{R}(a_1 \ldots a_k)$, которое отбирает подпоследовательность, состоящую из членов, следующих сразу за появлением $a_1 \ldots a_k$, определяет бесконечную подпоследовательность, и если эта подпоследовательность 1-распределена, то каждый из наборов, состоящих из $k + 1$ элементов $a_1 \ldots a_k a_{k+1}$, при условии, что $0 \leq a_{k+1} < b$, появляется в $\langle X_n\rangle$ с вероятностью $1/b^{k+1}$. Таким образом, индукцией по k можно доказать, что последовательность, удовлетворяющая определению **R5**, k-распределена для всех k. Аналогичным образом, рассматривая "композицию" правил построения подпоследовательностей (если $\mathcal{R}_1$ определяет бесконечную подпоследовательность $\langle X_n\rangle\,\mathcal{R}_1$, можно определить $\mathcal{R}_1\mathcal{R}_2$ как правило построения подпоследовательности, такое, что $\langle X_n\rangle\,\mathcal{R}_1\mathcal{R}_2 = ((\langle X_n\rangle\,\mathcal{R}_1)\mathcal{R}_2)$, мы приходим к выводу, что все подпоследовательности, описанные в определении **R5**, ∞-распределены **(**см. упр. 32**)**.

Поскольку ∞-распределенность следует из определения **R5** как очень частный случай, можно надеяться на то, что мы, наконец, сформулировали искомое определение случайности. Но, увы, остается еще одна проблема! Вовсе не очевидно, что последовательности, удовлетворяющие определению **R4**, должны также удовлетворять определению **R5**. "Вычислимые правила построения

подпоследовательностей", которые мы ввели, всегда перечисляют подпоследовательности $\langle X_{s_n}\rangle$, для которых $s_0 < s_1 < \ldots$, и это приводит к так называемым "рекурсивным последовательностям". Они не включают в себя все рекурсивно перечислимые последовательности, допускаемые определением R4, где последовательность $\langle s_n\rangle$ не обязана быть монотонной; должно лишь соблюдаться условие $s_n \neq s_m$, при $n = m$.

Столкнувшись с таким препятствием, мы приходим к комбинации определений R4 и R5.

Определение R6. *b-ичная последовательность $\langle X_n\rangle$ называется "случайной", если для любого эффективного алгоритма, определяющего бесконечную последовательность различных неотрицательных целых чисел $\langle s_n\rangle$ как функцию от n и значений $X_{s_0}, \ldots, X_{s_{n-1}}$, соответствующая подпоследовательность $\langle X_{s_n}\rangle$ "случайна" в смысле определения R5.*

Последовательность $\langle U_n\rangle$ на $[0,1)$ называется "случайной", если b-ичная последовательность $\langle \lfloor bU_n\rfloor\rangle$ "случайна" при всех целых числах $b \geq 2$.

Автор утверждает, что это определение, несомненно, отвечает всем разумным философским требованиям, предъявляемым к понятию случайности, и, таким образом, дает ответ на главный вопрос, поставленный здесь.

D. Существование случайных последовательностей. Мы видели, что определение R3 потому оказалось слишком сильным, что ни одна последовательность ему не удовлетворяла, и вводя определения R4, R5 и R6, мы старались сохранить основные свойства определения R3. Для того чтобы показать, что определение R6 не является слишком ограничительным, следует доказать факт существования последовательностей, удовлетворяющих соответствующим требованиям. Исходя из интуитивных соображений, никто не сомневается в их существовании, поскольку каждый верит в то, что истинно случайные последовательности существуют и удовлетворяют определению R6. Однако чтобы убедиться в состоятельности определения, необходимо доказательство.

Интересный метод построения последовательностей, удовлетворяющих определению R5, предложил А. Вальд. Сначала строится очень простая 1-распределенная последовательность.

Лемма T. *Пусть в двоичной системе счисления определена последовательность действительных чисел $\langle V_n\rangle$:*

$$V_0 = 0, V_1 = .1, V_2 = .01, V_3 = .11, V_4 = .001, \ldots, V_n = .c_r \ldots c_1 1, \qquad \text{если } n = 2^r + c_1 2^{r-1} + \cdots + c_r. \quad (29)$$

Пусть $I_{b_1\ldots b_r}$ обозначает множество таких действительных чисел на отрезке $[0,1)$, двоичное представление которых начинается с $0.b_1 \ldots b_r$. Таким образом,

$$I_{b_1\ldots b_r} = [0.b_1 \ldots b_r, 0.b_1 \ldots b_r + 2^{-r}). \quad (30)$$

Тогда, если $\nu(n)$ обозначает количество чисел V_k, содержащихся в $I_{b_1\ldots b_r}$ при $0 \leq k < n$, имеет место неравенство

$$\left|\nu(n)/n - 2^{-r}\right| \leq 1/n. \quad (31)$$

Доказательство. Поскольку $\nu(n)$ есть число тех k, для которых $k \bmod 2^r = b_r \ldots b_1$, мы имеем $\nu(n) = t$ или $t+1$, когда $\lfloor n/2^r\rfloor = t$. Таким образом, $|\nu(n) - n/2^r| \leq 1$. ∎

Из формулы (31) следует, что последовательность $\langle \lfloor 2^r V_n\rfloor\rangle$ является равномерно распределенной 2^r-ичной последовательностью; отсюда и из теоремы А заключаем, что $\langle V_n\rangle$—равномерно распределенная на $[0,1)$ последовательность. В самом деле, ясно, что $\langle V_n\rangle$ настолько равномерно распределена, насколько может быть равномерно распределенной последовательность на $[0,1)$! (Дальнейшее обсуждение свойств этой и связанных с ней последовательностей имеется в статьях И. ван дер Корпута (*Proc. Koninklijke Nederlandse Akademie van Wetenschappen,* **38** (1935), 813–821, 1058–1066) и Дж. Холтона (*Numerische Mathematik,* **2** (1960), 84–90, 196).

Пусть теперь $\mathcal{R}_1, \mathcal{R}_2, \ldots$—бесконечное множество правил построения подпоследовательностей. Мы хотим найти последовательность $\langle U_n\rangle$, все бесконечные подпоследовательности $\langle U_n\rangle\,\mathcal{R}_j$ которой равномерно распределены.

Алгоритм W. *(Последовательность Вальда.)* Эта процедура определяет последовательность $\langle U_n\rangle$ на $[0,1)$, если задано бесконечное множество правил построения подпоследовательностей $\mathcal{R}_1, \mathcal{R}_2, \ldots$, определяющих подпоследовательности последовательностей *рациональных* чисел на $[0,1)$. При вычислении требуется бесконечно большое количество вспомогательных переменных $C[a_1, \ldots, a_r]$, где $r \geq 1$ и $a_j = 0$ или 1, $1 \leq j \leq r$. В начальный момент времени все эти переменные равны нулю.

W1 **[Начальная установка n.]** Установить $n \leftarrow 0$.
W2 **[Начальная установка r.]** Установить $r \leftarrow 1$.
W3 **[Проверить $\mathcal{R}_r$.]** Если элемент U_n, должен попасть в подпоследовательность, определяемую $\mathcal{R}_r$, на основании значений U_k, где $0 \leq k < n$, то нужно присвоить $a_r \leftarrow 1$, в противном случае—присвоить $a_r \leftarrow 0$.
W4 **[$B[a_1, \ldots, a_r]$ полно?]** Если $C[a_1, \ldots, a_r] < 3 \cdot 4^{r-1}$, перейти к **W6**.
W5 **[Увеличить r.]** Установить $r \leftarrow r + 1$ и возвратиться к **W3**.
W6 **[Установить U_n.]** Увеличить $C[a_1, \ldots, a_r]$ на 1. Пусть k есть новое значение $C[a_1, \ldots, a_r]$. Установить $U_n \leftarrow V_k$, где величина V_k определена выше в лемме **T**.
W7 **[Увеличить n.]** Увеличить n на 1 и возвратиться к **W2**. ∎

Строго говоря, это не есть алгоритм, поскольку он не конечен. Легко, однако, изменить его так, чтобы он заканчивался, когда n достигает заданной величины. Читателю легче будет почувствовать идею приведенного построения, если он попробует "прокрутить" его вручную, заменив при этом число $3 \cdot 4^{r-1}$ на шаге **W4** на 2^r.

Алгоритм **W** не предназначен для применения в качестве датчика случайных чисел, он служит лишь теоретическим целям.

Теорема W. *Пусть U_n—последовательность рациональных чисел, определенная с помощью алгоритма W, и k—положительное целое число. Если подпоследовательность $\langle U_n \rangle \mathcal{R}_k$ бесконечна, то она* 1*-распределена.*

Доказательство. Пусть $A[a_1, \ldots, a_r]$ обозначает подпоследовательность (может быть, пустую) последовательности $\langle U_n \rangle$, содержащую те и только те элементы U_n, которые для всех j, таких, что $1 \leq j \leq r$, принадлежат подпоследовательности $\langle U_n \rangle \mathcal{R}_j$, если $a_j = 1$, и не принадлежат подпоследовательности $\langle U_n \rangle \mathcal{R}_j$, если $a_j = 0$.

Достаточно доказать, что для всех $r \geq 1$ и всех пар двоичных чисел $a_1 \ldots a_r$ и $b_1 \ldots b_r$ имеет место равенство $\Pr(U_n \in I_{b_1 \ldots b_r}) = 2^{-r}$ по отношению к подпоследовательности $A[a_1 \ldots a_r]$ в том случае, когда последняя бесконечна **[см. (30)]**. В самом деле, если $r \geq k$, то бесконечная последовательность $\langle U_n \rangle \mathcal{R}_k$ представляет собой конечное объединение непересекающихся подпоследовательностей $A[a_1, \ldots, a_r]$ для $a_k = 1$ и $a_j = 0$ или 1 при $1 \leq j \leq r$, $j \neq k$; следовательно, $\Pr(U_n \in I_{b_1 \ldots b_r}) = 2^{-r}$ по отношению к $\langle U_n \rangle \mathcal{R}_k$ **(см. упр. 33)**. Достаточно воспользоваться еще теоремой **A**, чтобы показать, что последовательность 1-распределена.

Пусть $B[a_1, \ldots, a_r]$ обозначает подпоследовательность элементов $\langle U_n \rangle$, для которых $C[a_1, \ldots, a_r]$ увеличивается на единицу на шаге **W6** алгоритма. Как видно из алгоритма, $B[a_1, \ldots, a_r]$—конечная последовательность, максимальное число элементов которой равно $3 \cdot 4^{r-1}$. Все члены $A[a_1, \ldots, a_r]$, кроме конечного их числа, берутся из подпоследовательностей $B[a_1, \ldots, a_r, \ldots, a_t]$, где $a_j = 0$ или 1 при $r < j \leq t$.

Предположим теперь, что $A[a_1, \ldots, a_r]$ бесконечна и $A[a_1, \ldots, a_r] = \langle U_{s_n} \rangle$, где $s_0 < s_1 < s_2 < \ldots$. Если N—большое целое число, такое, что $4^r \leq 4^q < N \leq 4^{q+1}$, количество значений $k < N$, при которых U_{s_k} содержится в $I_{b_1 \ldots b_r}$, равно (за исключением конечного числа элементов в начале подпоследовательности)

$$\nu(N) = \nu(N_1) + \cdots + \nu(N_m).$$

Здесь m—количество подпоследовательностей $B[a_1, \ldots, a_t]$, перечисленных выше, в которых U_s появляется при некотором $k < N$, N_j—количество значений k, при которых U_{s_k} находится в соответствующей подпоследовательности, и $\nu(N_j)$—количество таких значений, которые также находятся в $I_{b_1 \ldots b_r}$. Таким образом, из леммы **T** следует, что

$$\begin{aligned} \left|\nu(N) - 2^{-r}N\right| &= \left|\nu(N_1) - 2^{-r}N_1 + \cdots + \nu(N_m) - 2^{-r}N_m\right| \leq \\ &\leq \left|\nu(N_1) - 2^{-r}N_1\right| + \cdots + \left|\nu(N_m) - 2^{-r}N_m\right| \leq \\ &\leq m \leq 1 + 2 + 4 + \cdots + 2^{q-r+1} < 2^{q+1}. \end{aligned}$$

Неравенство относительно m следует из того факта, что U_{s_N} при нашем выборе N находится в $B[a_1, \ldots, a_t]$ для некоторого $t \leq q + 1$.

Таким образом,

$$\left|\nu(N)/N - 2^{-r}\right| \leq 2^{q+1}/N < 2/\sqrt{N}.$$

∎

Для того чтобы наконец показать существование последовательностей, удовлетворяющих определению **R5**, заметим прежде всего, что если $\langle U_n \rangle$—последовательность рациональных чисел на $[0, 1)$

и если $\mathcal{R}$—вычислимое правило построения подпоследовательности членов b-ичной последовательности, мы можем превратить $\mathcal{R}$ в вычислимое правило $\mathcal{R}'$ построения подпоследовательности членов $\langle U_n\rangle$, положив $f'_n(x_1, \ldots, x_n)$ в $\mathcal{R}'$ равным $f_n(\lfloor bx_1\rfloor, \ldots, \lfloor bx_n\rfloor)$ в $\mathcal{R}$. Если последовательность $\langle U_n\rangle\,\mathcal{R}'$ равномерно распределена, то этим же свойством обладает и $\langle\lfloor bU_n\rfloor\rangle\,\mathcal{R}$. Множество всех вычислимых правил построения подпоследовательностей b-ичных последовательностей при всех значениях b счетно (поскольку существует лишь счетное количество эффективных алгоритмов), поэтому его элементы можно перечислить в виде некоторой последовательности $\mathcal{R}_1$, $\mathcal{R}_2$, ... Отсюда следует, что алгоритм W определяет последовательность на $[0, 1)$, которая является случайной в смысле определения R5.

Теперь мы оказались в несколько парадоксальном положении. Как отмечалось раньше, эффективный алгоритм, который определял бы последовательность, удовлетворяющую определению R4, существовать не может, и по той же причине не может существовать эффективный алгоритм, определяющий последовательность, удовлетворяющую определению R5. Доказательство существования такой случайной последовательности по необходимости должно быть неконструктивным. Каким же образом эта последовательность получается по алгоритму W?

Противоречия здесь нет. Дело в том, что невозможно с помощью эффективного алгоритма пронумеровать множество всех алгоритмов. Другими словами, эффективный алгоритм, который выбирал бы j-е вычислимое правило построения подпоследовательности $\mathcal{R}_j$, не может существовать; это следует из того, что не может существовать эффективный алгоритм, который определял бы, состоит ли данный вычислительный метод из конечного числа шагов. (Мы вернемся к этому вопросу в гл. 11.) Однако важные большие классы алгоритмов *могут* быть систематически перечислены. Например, из построения алгоритма W видно, что с помощью эффективного алгоритма можно построить последовательность, удовлетворяющую определению R5, если мы ограничимся "примитивно рекурсивными" правилами построения подпоследовательностей.

Если видоизменить шаг W6 алгоритма W так, чтобы там происходила установка $U_n \leftarrow V_{k+t}$ (а не V_k), где t—любое неотрицательное целое число, зависящее от a_1, ..., a_r, то можно показать, что существует *несчетное* множество последовательностей на $[0, 1)$, удовлетворяющих определению R5.

Другой, менее прямой, путь доказательства существования несчетного множества случайных последовательностей, основанный на теории меры, для последовательностей, удовлетворяющих даже сильному определению R6, дает

Теорема М. *Пусть действительное число x, $0 \le x < 1$, поставлено в соответствие двоичной последовательности $\langle X_n\rangle$ таким образом, что двоичное представление x есть $0.X_0X_1\ldots$. Имея в виду это соответствие, можно утверждать, что почти все x соответствуют двоичным последовательностям, которые являются случайными в смысле определения R6. (Другими словами, множество тех действительных x, которые соответствуют неслучайным в смысле определения R6 последовательностям, имеет меру нуль.)*

Доказательство. Пусть $\mathcal{S}$—эффективный алгоритм, определяющий бесконечную последовательность различных неотрицательных целых чисел $\langle s_n\rangle$, такой, что s_n зависит только от n и X_{s_k}, где $0 \le k < n$, и $\mathcal{R}$—вычислимое правило построения подпоследовательностей. Тогда из любой двоичной последовательности $\langle X_n\rangle$ можно получить подпоследовательность $\langle X_{s_n}\rangle\,\mathcal{R}$, и определение R6 утверждает, что эта подпоследовательность должна быть или конечной, или 1-распределенной. Достаточно доказать, что *при заданных $\mathcal{R}$ и $\mathcal{S}$ множество $N(\mathcal{R},\mathcal{S})$ действительных чисел x, соответствующих $\langle X_n\rangle$ и таких, что последовательность $\langle X_{s_n}\rangle\,\mathcal{R}$ бесконечна и не является 1-распределенной, имеет меру нуль.* В самом деле, x имеет неслучайное двоичное представление в том и только том случае, когда x принадлежит объединению $\cup N(\mathcal{R},\mathcal{S})$, просуммированному по счетному множеству $\mathcal{R}$ и $\mathcal{S}$.

Пусть, таким образом, $\mathcal{R}$ и $\mathcal{S}$ заданы. Рассмотрим множество $T(a_1a_2\ldots a_r)$, которое определяется для любых двоичных чисел $a_1a_2\ldots a_r$ как множество всех тех x, соответствующих $\langle X_n\rangle$, что $\langle X_{s_n}\rangle\,\mathcal{R}$ имеет $\ge r$ элементов, причем первые r элементов равны соответственно a_1, a_2, ..., a_r. Сначала мы докажем, что

$$\text{мера множества } T(a_1a_2\ldots a_r) \le 2^{-r}. \tag{32}$$

Заметим вначале, что множество $T(a_1a_2\ldots a_r)$ измеримо: каждый элемент из $T(a_1a_2\ldots a_r)$ есть действительное число $x = 0.X_0X_1\ldots$, для которого существует целое число m, такое, что алгоритм $\mathcal{S}$ определяет различные значения s_0, s_1, ..., s_m, и правило $\mathcal{S}$ определяет подпоследовательность X_{s_0}, X_{s_1}, ..., X_{s_m}, такую, что X_{s_m} есть r-й элемент этой последовательности. Множество всех действительных чисел $y = 0.Y_0Y_1\ldots$, таких, что $Y_{s_k} = X_{s_k}$ при $0 \le k \le m$, также принадлежит множеству $T(a_1a_2\ldots a_r)$ и является измеримым множеством, состоящим из конечного объединения двоичных подынтервалов $I_{b_1\ldots b_t}$. Поскольку множество таких интервалов счетно, $T(a_1a_2\ldots a_r)$ является

счетным объединением двоичных интервалов, и, следовательно, оно измеримо. Более того, из этого рассуждения следует, что мера $T(a_1 \dots a_{r-1}0)$ равна мере $T(a_1 \dots a_{r-1}1)$, поскольку последнее множество есть объединение двоичных интервалов, полученных из предшествующего при дополнительном требовании, что $Y_{s_k} = X_{s_k}$ для $0 \le k < m$ и $Y_{s_m} \ne X_{s_m}$. Поскольку

$$T(a_1 \dots a_{r-1}0) \cup T(a_1 \dots a_{r-1}1) \subseteq T(a_1 \dots a_{r-1}),$$

мера множества $T(a_1 a_2 \dots a_r)$ не превосходит половины меры множества $T(a_1 \dots a_{r-1})$. Неравенство (32) получается индукцией по r.

Теперь, когда справедливость неравенства (32) установлена, осталось доказать в основном следующее: двоичные представления почти всех действительных чисел равномерно распределены. Далее в нескольких абзацах, где иллюстрируется типичная в математическом анализе техника получения оценок, представлено довольно длинное, но не трудное доказательство этого факта.

Пусть $0 < \varepsilon < 1$ и $B(r, \varepsilon)$ есть $\bigcup T(a_1 \dots a_r)$, где объединение берется по всем двоичным числам $a_1 \dots a_r$, таким, что количество $\nu(r)$ нулей среди $a_1, \dots, a_r$ удовлетворяет неравенству

$$\left| \nu(r) - \frac{1}{2} r \right| \ge 1 + \varepsilon r.$$

Количество таких двоичных чисел равно $C(r, \varepsilon) = \sum \binom{r}{k}$, где суммирование ведется по значениям k, таким, что $\left|k - \frac{1}{2} r\right| \ge 1 + \varepsilon r$. Пусть $r = 2t$ есть целое четное число. Мы можем дать грубую оценку величины $\sum \binom{r}{k}$. Если $k > 0$, то

$$\binom{2t}{t+k} = \binom{2t}{t} \frac{t}{t+1} \frac{t-1}{t+2} \dots \frac{t-k+1}{t+k} < \binom{2t}{t} \frac{t}{t} \frac{t-1}{t} \dots \frac{t-k+1}{t} \le$$
$$\le \binom{2t}{t} e^{-0/t} e^{-1/t} \dots e^{-(k-1)/t} = \binom{2t}{t} e^{-k(k-1)/r}.$$

Таким образом,

$$C(r, \varepsilon) = 2 \sum_{k \ge 1 + \varepsilon r} \binom{2t}{t+k} \le 2 \binom{2t}{t} \sum_{k \ge 1 + \varepsilon r} e^{-k(k-1)/r} \le$$
$$\le 2 \binom{2t}{t} t e^{-(1+\varepsilon r)\varepsilon)} < r \binom{r}{t} e^{-\varepsilon^2 r}.$$

Аналогично, для $r = 2t + 1$ получаем

$$C(r, \varepsilon) < r \binom{r}{t} e^{-\varepsilon^2 r}.$$

Используя теперь (32), выводим

$$\text{мера множества } B(r, \varepsilon) \le 2^{-r} C(r, \varepsilon) < r e^{-\varepsilon^2 r}. \tag{33}$$

Далее определим множество

$$B^*(r, \varepsilon) = B(r, \varepsilon) \cup B(r+1, \varepsilon) \cup B(r+2, \varepsilon) \cup \dots .$$

Мера $B^*(r, \varepsilon)$ не превосходит $\sum_{k \ge r} k e^{-\varepsilon^2 k}$. Последняя величина является остатком сходящегося ряда, поэтому

$$\lim_{r \to \infty} (\text{меры } B^*(r, \varepsilon)) = 0. \tag{34}$$

Если теперь x—действительное число, такое, что его двоичное представление $0.X_0X_1 \dots$ приводит к бесконечной последовательности $\langle X_{s_n} \rangle \mathcal{R}$, которая не является 1-распределенной, а $\nu(r)$ обозначает число нулей в первых r ее элементах, то

$$\left| \nu(r)/r - \frac{1}{2} \right| \ge 2\varepsilon$$

для некоторого $\varepsilon > 0$ и бесконечно многих r. Это значит, что x при всех r содержится в $B^*(r, \varepsilon)$. Таким образом, окончательно находим, что

$$N(\mathcal{R}, \mathcal{S}) = \bigcup_{t \ge 2} \bigcap_{r \ge 1} B^*(r, 1/t).$$

Из формулы (34) следует, что $\bigcap_{r\geq 1} B^*(r, 1/t)$ при всех t имеет меру нуль; следовательно, $N(\mathcal{R}, \mathcal{S})$ также имеет меру нуль. ∎

Из существования *двоичных* последовательностей, удовлетворяющих определению R6, следует существование последовательностей на $[0, 1)$, случайных в смысле этого определения. См. по этому поводу упр. 36. Тем самым мы установили состоятельность определения R6.

E. Случайные конечные последовательности. Выше приводилось соображение о том, что понятие случайности для конечных последовательностей ввести невозможно, поскольку всякая заданная конечная последовательность ничуть не хуже любой другой. Несмотря на это, почти каждый согласится с тем, что последовательность 011101001 "более случайна", чем последовательность 101010101, а последняя "более случайна", чем 000000000. Хотя справедливо утверждение, что истинно случайная последовательность локально может быть неслучайной, мы бы предпочли обнаружить такую неслучайность только в длинной, а не в короткой конечной последовательности.

Существует несколько подходов к определению случайности конечной последовательности, и мы наметим лишь несколько относящихся сюда идей. Будем рассматривать только b-ичные последовательности.

Если задана b-ичная последовательность $X_1, X_2, \ldots, X_N$, то можно сказать, что

$$\Pr(S(n)) \approx p, \qquad \text{если } |\nu(N)/N - p| \leq 1/\sqrt{N},$$

где $\nu(n)$—величина, введенная в определении A в начале настоящего параграфа. Приведенную выше последовательность можно назвать "k-распределенной", если

$$\Pr(X_n X_{n+1} \ldots X_{n+k-1} = x_1 x_2 \ldots x_k) \approx 1/b^k$$

для всех b-ичных чисел $x_1 x_2 \ldots x_k$. (Ср. с определением D. К сожалению, по этому новому определению последовательность может быть k-распределена, даже если она не является $(k-1)$-распределенной.)

Теперь можно ввести понятие случайности аналогично тому, как это было сделано в определении R1.

Определение Q1. *b-ичная последовательность длины N называется "случайной", если она k-распределена (в указанном выше смысле) при всех положительных целых k, таких, что $k \leq \log_b N$.*

В соответствии с этим определением, имеются, например 170 неслучайных двоичных последовательностей длины 11:

00000001111	10000000111	11000000011	11100000001
00000001110	10000000110	11000000010	11100000000
00000001101	10000000101	11000000001	10100000001
00000001011	10000000011	01000000011	01100000001
00000000111			

плюс 01010101010 и все последовательности, в которых имеется не менее девяти нулей, плюс все последовательности, полученные из предшествующих взаимной заменой нулей и единиц.

Подобным же образом можно ввести определение, аналогичное определению R6, для конечных последовательностей. Пусть A есть множество алгоритмов, каждый из которых представляет собой процедуру получения подпоследовательности $\langle X_{s_n} \rangle \mathcal{R}$, аналогично тому, как это сделано при доказательстве теоремы M.

Определение Q2. *b-ичная последовательность $X_1, X_2, \ldots, X_N$ называется (n, ε)-случайной по отношению к множеству A алгоритмов, если для каждой подпоследовательности $X_{t_1}, X_{t_2}, \ldots, X_{t_m}$, определенной с помощью алгоритма, принадлежащего множеству A, справедливо либо неравенство $m < n$, либо неравенство*

$$\left| \frac{1}{m} \nu_a(X_{t_1}, \ldots, X_{t_m}) - \frac{1}{b} \right| \leq \varepsilon \qquad \textit{при } 0 \leq a < b.$$

Здесь $\nu_a(x_1, \ldots, x_m)$ обозначает число появлений a в последовательности $x_1, \ldots, x_m$.

(Другими словами, каждая достаточно длинная подпоследовательность, определенная с помощью алгоритма, принадлежащего множеству A, должна быть приближенно равномерно распределена.) Основная идея состоит в том, чтобы составить множество A из "простых" алгоритмов, число же их (и сложность) может возрастать с ростом N.

В качестве примера рассмотрим двоичные последовательности, и пусть A состоит из четырех алгоритмов.

(a) Взять всю последовательность.
(b) Взять члены последовательности через один, начиная с первого.
(c) Взять члены последовательности, стоящие после нуля.
(d) Взять члены последовательности, стоящие после единицы.

Так, последовательность $X_1, \ldots, X_8$ является $(4, 1/8)$-случайной, если

по алгоритму (a): $\left|\frac{1}{8}(X_1 + \cdots + X_8) - \frac{1}{2}\right| \le \frac{1}{8}$, т. е. если в ней имеется 3,4 или 5 единиц;
по алгоритму (b): $\left|\frac{1}{4}(X_1 + X_3 + X_5 + X_7) - \frac{1}{2}\right| \le \frac{1}{8}$, т. е. если в ней имеются две единицы на местах с нечетными номерами;
по алгоритму (c): нужно рассмотреть три возможности в зависимости от того, сколько нулей стоят на местах $X_1, \ldots, X_7$: если там стоят 2 или 3 нуля, то проверять ничего не нужно (поскольку $n = 4$), если там 4 нуля, за ними должны стоять два нуля и две единицы, и, наконец, если там 5 нулей, за ними должны стоять два или три нуля.
по алгоритму (d): условия, аналогичные полученным по алгоритму (c).

Итак, только следующие двоичные последовательности длины 8 являются по этим правилам $\left(4, \frac{1}{8}\right)$-случайными:

00001011	00101001	01001110	01101000
00011010	00101100	01011011	01101100
00011011	00110010	01011110	01101101
00100011	00110011	01100010	01110010
00100110	00110110	01100011	01110110
00100111	00111001	01100110	

плюс последовательности, полученные из них взаимной заменой 0 и 1.

Ясно, что множество алгоритмов можно взять настолько большим, что ни одна последовательность не будет удовлетворять определению в случае, когда n и ε разумным образом малы. А. Н. Колмогоров доказал, что (n, e)-случайная двоичная последовательность *всегда существует* при любом заданном N, если число алгоритмов в A не превосходит

$$\frac{1}{2}e^{2n\varepsilon^2(1-\varepsilon)}. \tag{35}$$

Этот результат недостаточно силен для того, чтобы показать, что существуют последовательности, удовлетворяющие определению Q1, однако такие последовательности могут быть эффективно построены с помощью процедуры Риса в упр. 3.2.2-21.

Другой интересный подход к определению случайности указал П. Мартин-Лёф (*Information and Control,* **9** (1966), 602–619). Пусть заданы конечная b-ичная последовательность $X_1, \ldots, X_N$ и $l(X_1, \ldots, X_N)$—длина кратчайшей программы, которая получает эту последовательность на машине Тьюринга. (Определение понятия машины Тьюринга см. в гл. 11; вместо этого понятия мы могли бы использовать некоторые классы эффективных алгоритмов наподобие определенных в упр. 1.1-8.) Тогда $l(X_1, \ldots, X_N)$ служит мерой "нерегулярности" последовательности, и мы можем отождествить это понятие со случайностью. Последовательности длины N, максимизирующие $l(X_1, \ldots, X_N)$, можно назвать случайными. (С точки зрения практического получения случайных чисел на машине это определение случайности конечно наихудшее, какое только можно вообразить!)

В основных чертах такое же определение случайности независимо и почти одновременно предложил Г. Чьятин (*JACM,* **16** (1969), 145–159). Интересно заметить, что, хотя это определение в отличие от других никак не использует свойство равномерной распределенности, Мартин-Лёф и Чьятин доказали, что случайные последовательности этого типа также имеют ожидаемые свойства равномерной распределенности. Мартин-Лёф показал, что такие последовательности в некотором смысле удовлетворяют *всем* вычислимым статистическим проверкам случайности.

F. Выводы, история вопроса и библиография. Мы имеем теперь несколько определений разной степени случайности последовательности.

Бесконечные ∞-распределенные последовательности обладают многими полезными свойствами, которыми должны обладать случайные последовательности, и, кроме того, для них существует развитая теория. (В упражнениях, помещенных ниже, вводятся несколько важных свойств ∞-распределенных последовательностей, которые не упоминались в тексте.) Таким образом, определение R1 подходит для теоретических исследований случайности.

Понятие ∞-распределенной b-ичной последовательности ввел в 1909 г. Эмиль Борель. Он предложил также понятие (m, k)-распределенной последовательности и показал, что b-ичные представления

почти всех действительных чисел (m,k)-распределены при всех значениях m и k. Он назвал такие числа *нормальными* по отношению к основанию b. Превосходное обсуждение этого вопроса есть в его хорошо известной книге "Leçons sur la théorie des fonctions" (2-е изд. 1914), 182–216. Более поздние результаты и соответствующая библиография, касающиеся нормальных чисел, есть в статье В. Шмидта (*Pacific Journal of Math.,* **10** (1960), 661–672).

Понятие ∞-распределенной последовательности *действительных* чисел, которую называют также *совершенно равномерно распределенной последовательностью,* ввел Дж. Фрэнклин в статье "Deterministic Simulation of Random Processes" [*Math. Comp.,* **17** (1963), 28–59)]. В этой важной работе решено несколько очень интересных задач, связанных с k-распределенными последовательностями, и исследованы свойства распределенности многих специальных последовательностей.

Мы видели, однако, что ∞-распределенные последовательности не обязаны быть настолько беспорядочными, чтобы их нужно было считать совершенно "случайными". Выше были сформулированы три определения R4, R5 и R6 для того, чтобы наложить на последовательности добавочные условия. Особенно подходящим определением бесконечной случайной последовательности представляется R6. Оно дает точную количественную формулировку, которая, по-видимому, совпадает с интуитивным понятием истинной случайности.

Если говорить об истории вопроса, развитие этих определений во многом обязано Р. фон Мизесу и его поискам хорошего определения "вероятности". Он (*Math. Zeitschrift,* **5** (1919), 52–99) предложил определение, по духу близкое определению R5, хотя по форме настолько сильное (наподобие нашего определения R3), что последовательности, удовлетворяющие наложенным условиям, не могут существовать. Многие заметили это обстоятельство, и А. Коуплэнд (*Amer. J. Math.,* **50** (1928), 535–552) предложил ослабить определение фон Мизеса, вводя так называемые "допустимые числа" (или последовательности Бернулли). Они эквивалентны ∞-распределенным на $[0,1)$ последовательностям, в которых все U_n заменяются на 1, если $U_n < p$, или на 0, если $U_n \geq p$, причем вероятность p задана. Таким образом, Коуплэнд по существу предлагал возвратиться к определению R1. А. Вальд показал, что столь радикально ослаблять определение фон Мизеса не нужно, а достаточно ввести счетное множество правил построения подпоследовательностей. В важной статье [*Ergebnisse eines math. Kolloquiums,* **8** (Vienna, 1937), 38–72] Вальд по существу доказал теорему W, хотя и сделал ошибочное утверждение, состоящее в том, что последовательность, построенная с помощью алгоритма W, удовлетворяет также более сильному требованию "мера $A = \Pr(U_n \in A)$" для всех измеримых по Лебегу $A \subseteq [0,1)$. Мы видели, что ни одна последовательность не может удовлетворить этому требованию.

Определение R4 впервые было сформулировано Д. Лавлэндом (*Zeit. fur Math. Logik und Grundlagen d. Math.,* **12** (1966), 279–294), который обнаружил, что при помощи алгоритма W можно доказать существование двоичной последовательности, удовлетворяющей определению R5, но не R4.

Когда Вальд писал свою статью, понятие "вычислимости" было лишь недавно создано, и А. Чёрч [*Bulletin AMS,*, **47** (1940), 130–135] показал, как в теорию Вальда можно ввести строгое понятие "эффективного алгоритма" для того, чтобы сделать его определения совершенно точными. Существенное продвижение к определению R6 сделал А. Н. Колмогоров [*Sankhyā,* series A, **25** (1963), 369–376], который в той же статье предложил определение Q2 для конечных последовательностей.

Дальнейшие связи между случайными последовательностями и теорией рекурсивных функций были выявлены Д. Лавлэндом [*Trans. AMS,* **125** (1966), 497–510]. См. также работу К.-П. Шнорра [*Z. Wahr. verw. Geb.,* **14** (1969), 27–35], где приведены сильные зависимости между случайными последовательностями и "категориями меры нуль", определенными Броуэром в 1919 г.

В статьях Чёрча и Колмогорова рассматривались только двоичные последовательности, для которых $\Pr(X_1 = 1) = p$ при заданной вероятности p. Тем самым рассуждения, проведенные в этой главе, являются несколько более общими, поскольку последовательность на $[0,1)$ представляет одновременно все p.

В другой статье [*Проблемы передачи информации,* **1**, (1965), 3–11] Колмогоров рассмотрел задачу об определении "количества информации" последовательности, и эта работа привела Мартин-Лёфа к интересному определению конечных случайных последовательностей через "нерегулярность" [*IEEE Trans.,* **IT-14** (1968), 662–664]. Еще одно определение случайности, которое можно было бы поместить между определениями Q1 и Q2, было сформулировано довольно давно А. С. Безиковичем [*Math. Zeitschrift,* **39** (1934), 146–156].

Дальнейшее обсуждение свойств случайных последовательностей можно найти в книге К. Поппера "The Logic of Scientific Discovery" (London, 1959). Особенно интересны результаты, приведенные на стр. 162–163, которые он впервые опубликовал в 1934 г.

Интересные упражнения, связанные с последовательностями на $[0,1)$, есть в книге Д. Полиа, Г. Сегё, "Задачи и теоремы из анализа", часть 1 (*ГИТТЛ,* Москва, 1956 г.); о равномерно распределен-

ных b-ичных последовательностях написано в работах А. Нивена (*Trans. Amer. Math. Soc.*, **98** (1961); 52–61) и Б. Зэйна (*AMM*, **71** (1964), 162–164).

Упражнения

1. [10]Может ли периодическая последовательность быть равномерно распределенной?
2. [10] Рассмотрите периодическую двоичную последовательность 0, 0, 1, 1, 0, 0, 1, 1, ... Является ли она 1-распределенной? 2-распределеиной? 3-распределенной?
3. [M22] Постройте 3-распределенную периодическую троичную последовательность.
>4. [BM22] Рассмотрите последовательность $U_n = (2^{\lfloor \log_2(n+1) \rfloor}/3) \bmod 1$, чему равна $\Pr(U_n < 1/2)$?
5. [BM14] Докажите, что для любых двух утверждений $S(n)$, $T(n)$ справедливо равенство $\Pr(S(n)$ е $T(n)) + \Pr(S(n)$ есе $T(n)) = \Pr(S(n)) + \Pr(T(n))$ при условии, что по крайней мере три из этих пределов существуют. [Если, например, последовательность 2-распределена, мы находим, что

$$\Pr(u_1 \le U_n < v_1 \text{ есе } u_2 \le U_{n+1} < v_2) = v_1 - u_1 + v_2 - u_2 - (v_1 - u_1)(v_2 - u_2).]$$

6. [BM23] Пусть $S_1(n)$, $S_2(n)$, ... —бесконечная последовательность утверждений, касающихся взаимно исключающих событий, т. е. утверждения $S_i(n)$ и $S_j(n)$ при $i \ne j$ не могут быть истинными одновременно. Предположим, что $\Pr(S_j(n))$ существует для каждого $j \ge 1$. Покажите, что $\underline{\Pr}(S_j(n)$ истинно для некоторого $j \ge 1) \ge \sum_{j \ge 1} \Pr(S_j(n))$, и приведите пример, показывающий, что равенство не обязательно имеет место.
7. [BM27] Пусть, как и в предыдущем упражнении, задано бесконечно большое число взаимно исключающих событий $S_{ij}(n)$, $i \ge 1$, $j \ge 1$, таких, что $\Pr(S_{ij}(n))$ существует. Предположим, что $S_{ij}(n)$ при всех $n > 0$ истинно для точно одной пары целых чисел i, j. Если $\sum_{i,j \ge 1} \Pr(S_{ij}(n)) = 1$, то можно ли отсюда заключить, что для всех $i \ge 1$ величина $\Pr(S_{ij}(n))$ истинно для некоторого $j \ge 1$) существует и равна $\sum_{j \ge 1} \Pr(S_{ij}(n))$?
8. [M15] Докажите утверждение (13).
9. [BM20] Докажите лемму E. [*Указание:* рассмотреть выражение $\sum_{1 \le j \le m} (y_{jn} - \alpha)^2$.]
>10. [BM22] Где при доказательстве теоремы C использовался тот факт, что q кратно m?
>11. [BM20] Примените теорему C, чтобы доказать, что если последовательность $\langle U_n \rangle$ ∞-распределена, то этим свойством обладает и подпоследовательность $\langle U_{2n} \rangle$.
12. [BM20] Покажите, что k-распределенная последовательность удовлетворяет тесту "наибольшее из k" в следующем смысле: $\Pr(u \le \max(U_n, U_{n+1}, \ldots, U_{n+k-1} < v) = v^k - u^k$.
>13. [BM27] Покажите, что ∞-распределенная на $[0, 1)$ последовательность удовлетворяет проверке интервалов в следующем смысле. Пусть $0 \le \alpha < \beta \le 1$ и $p = \beta - \alpha$; положим $f(0) = 0$, и $f(n)$ для $n \ge 1$ есть наименьшее целое $m > f(n-1)$, такое, что $\alpha \le U_m < \beta$. Тогда $\Pr(f(n) - f(n-1) = k) = p(1-p)^{k-1}$.
14. [BM25] Покажите, что ∞-распределенная последовательность удовлетворяет проверке на монотонность в следующем смысле. Если $f(0) = 1$ и $f(n)$ для $n \ge 1$ есть наименьшее целое число $m > f(n-1)$, такое, что $U_{m-1} > U_m$, то

$$\Pr(f(n) - f(n-1) = k) = 2k/(k+1)! - 2(k+1)/(k+2)!.$$

>15. [BM30] Покажите, что ∞-распределенная последовательность удовлетворяет тесту собирателя купонов для случая, когда имеются только два сорта купонов, в следующем смысле. Пусть X_1, X_2, ... есть ∞-распределенная двоичная последовательность. Положим $f(0) = 0$ и пусть $f(n)$ для $n \ge 1$ есть наименьшее целое $m > f(n-1)$, такое, что $\{ X_{f(n-1)+1}, \ldots, X_m \}$ есть множество $\{ 0, 1 \}$. Докажите,что $\Pr(f(n) - f(n-1) = k) = 2^{1-k}$; $k \ge 2$. (Ср. с упр. 7.)
16. [BM38] Справедлив ли тест собирателя купонов для ∞-распределенных последовательностей в случае, когда имеется больше двух сортов купонов? (Ср. с предыдущим упражнением.)
17. [BM50] Франклин доказал, что если r — заданное рациональное число, то последовательность $U_n = (r^n \bmod 1)$ не является 2-распределенной. Существует ли рациональное число r, такое, что эта последовательность будет равномерно распределена? В частности, будет ли эта последовательность равномерно распределенной в случае, когда $r = 3/2$? [Ср. со статьей К. Малера (*Mathematika*, **4** (1957), 122–124).]
>18. [BM22] Докажите, что если U_0, U_1, ... k-распределена, то этим же свойством обладает и последовательность V_0, V_1, ..., где $V_n = \lfloor nU_n \rfloor / n$.
19. [BM46] Рассмотрите определение R4, в котором слово "∞-распределенной" заменено на "1-распределенной". Существует ли последовательность, которая удовлетворяет этому более слабому определению, но не является ∞-распределенной? (То есть является ли это определение действительно более слабым?)

20. **[BM50]** Удовлетворяет ли последовательность $U_n = (\theta^n \bmod 1)$ определению **R4** для почти всех действительных чисел $\theta > 1$? **(**Ответ на этот вопрос можно получить или в том случае, когда найден отрицательный ответ на вопрос упр. **19**, или в том случае, когда показано, что для любой последовательности различных положительных целых чисел s_0, s_1, s_2, ... последовательность $U_n = (\theta^{s_n} \bmod 1)$ ∞-распределена для почти всех $\theta > 1$.**)**
21. **[BM20]** Пусть S—множество и $\mathcal{M}$—совокупность его подмножеств. Предположим, что p — функция, заданная на множествах из $\mathcal{M}$, принимающая на них действительные значения, и что $p(M)$ обозначает вероятность того, что "случайно" выбранный элемент из S принадлежит M. Обобщите определения **B** и **D** так, чтобы получить хорошее определение понятия k-распределенной последовательности $\langle Z_n \rangle$ элементов множества S по отношению к распределению вероятностей p.

>22. **[BM30]** (Герман Вейль.) Покажите, что последовательность $\langle U_n \rangle$ k-распределена на $[0, 1)$ в том и только том случае, когда

$$\lim_{N\to\infty} \frac{1}{N} \sum_{0\le n<N} \exp(2\pi i(c_1U_n + \cdots + c_kU_{n+k-1})) = 0$$

для каждого множества целых чисел c_1, c_2, ..., c_k, где не все c_j одновременно равны нулю.

23. **[M34]** Покажите, что b-ичная последовательность $\langle X_n \rangle$ k-распределена в том и только том случае, когда все последовательности $\langle c_1X_n + c_2X_{n+1} + \cdots + c_kX_{n+k-1} \rangle$ *равномерно распределены* при условии, что c_1, c_2, ..., c_k являются целыми числами с нод$(c_1, \ldots, c_k) = 1$.
24. **[M30]** Покажите, что последовательность $\langle U_n \rangle$ k-распределена на $[0, 1)$ в том и только том случае, когда все последовательности $\langle c_1U_n + c_2U_{n+1} + \cdots + c_kU_{n+k-1} \rangle$ *равномерно распределены* при условии, что c_1, c_2, ..., c_k—некоторые целые числа, не равные нулю одновременно.
25. **[BM20]** Последовательность называется "белой последовательностью", если все коэффициенты последовательной корреляции равны нулю, т. е. если соотношение из следствия **S** справедливо при *всех* $k \ge 1$. **(**Как видно из следствия **S**, ∞-распределенная последовательность является белой.**)** Покажите, что если $[0, 1)$-последовательность равномерно распределена, то она является белой тогда и только тогда, когда

$$\lim_{n\to\infty} \frac{1}{n} \sum_{0\le j<n} (U_j - 1/2)(U_{j+k} - 1/2) = 0 \qquad \text{при всех } k \ge 1.$$

26. **[BM34]** (Дж. Фрэнклин.) Определенная в предыдущем упражнении белая последовательность, безусловно, может не быть случайной. Пусть U_0, U_1, ... есть ∞-распределенная последовательность. Определим последовательность V_0, V_1, ... следующим образом:

$$(V_{2n-1}, V_{2n}) = (U_{2n-1}, U_{2n}), \qquad \text{если } (U_{2n-1}, U_{2n}) \in G,$$
$$(V_{2n-1}, V_{2n}) = (U_{2n}, U_{2n-1}), \qquad \text{если } (U_{2n-1}, U_{2n}) \notin G,$$

где G—множество $\{\, (x, y) \mid x - (1/2) \le y \le x$ **ece** $x + (1/2) \le y \,\}$. Покажите, что **(a)** V_0, V_1, ... есть равномерно распределенная и белая последовательность, **(b)** $\Pr(V_n > V_{n+1}) = 5/8$. **(**Отсюда видна слабость теста последовательной корреляции.**)**
27. **[BM49]** Пусть $\langle V_n \rangle$ — равномерно распределенная белая последовательность. Является ли число $5/8$ в предыдущем упражнении наибольшим возможным значением величины $\Pr(V_n > V_{n+1})$?

>28. **[M24]** Используя последовательность **(11)**, постройте 3-распределенную на $[0, 1)$ последовательность, для которой $\Pr(U_{2n} \ge 1/2) = 3/4$.
29. **[BM34]** Пусть X_0, X_1, ... есть $(2k)$-распределенная двоичная последовательность. Покажите, что

$$\overline{\Pr}(X_{2n} = 0) \le 1/2 + \binom{2k-1}{k} \Big/ 2^{2k}.$$

>30. **[M39]** Постройте $(2k)$-распределенную двоичную последовательность, для которой справедливо равенство

$$\Pr(X_{2n} = 0) = 1/2 + \binom{2k-1}{k} \Big/ 2^{2k}.$$

(Отсюда следует, что неравенство в предыдущем упражнении нельзя улучшить.**)**
31. **[M30]** Покажите, что существует последовательность на $[0, 1)$ удовлетворяющая определению **R5**, но такая, что $\nu_n/n \ge 1/2$ для всех $n > 0$, где ν_n есть число тех $j < n$, для которых $U_j < 1/2$. **(**Это можно рассматривать как неслучайное свойство последовательности.**)**

32. **[M24]** Пусть $\langle X_n\rangle$—"случайная" b-ичная последовательность в смысле определения **R5** и $\mathcal{R}$—вычислимое правило построения подпоследовательности, которое определяет бесконечную подпоследовательность $\langle X_n\rangle\,\mathcal{R}$. Покажите, что эта подпоследовательность не только 1-распределена, но и "случайна" в смысле определения **R5**.
33. **[BM24]** Пусть $\langle U_{r_n}\rangle$ и $\langle U_{s_n}\rangle$—бесконечные, не имеющие общих элементов подпоследовательности последовательности $\langle U_n\rangle$. (Это значит, что $r_0 < r_1 < r_2 < \ldots$ и $s_0 < s_1 < s_2 < \ldots$—две возрастающие последовательности целых чисел и $r_m \neq s_n$ при любых m, n). Пусть $\langle U_{t_n}\rangle$ есть составная подпоследовательность, для которой $t_0 < t_1 < t_2 < \ldots$ и множество $\{\,t_n\,\} = \{\,r_n\,\} \cup \{\,s_n\,\}$. Покажите, что если $\Pr(U_{r_n} \in A) = \Pr(U_{s_n} \in A) = p$, то $\Pr(U_{t_n} \in A) = p$.
>34. **[M25]** Определите правила $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$ ... построения подпоследовательностей, такие, чтобы эти правила можно было бы использовать с алгоритмом W для получения эффективного алгоритма построения последовательности на $[0, 1)$, удовлетворяющей определению **R1**.
35. **[M50]** Обобщите, если это возможно, алгоритм W так, чтобы получались последовательности, удовлетворяющие более строгим условиям определения **R6**.
36. **[BM30]** Пусть $\langle X_n\rangle$—двоичная последовательность, которая является "случайной" в смысле определения **R6**. Покажите, что последовательность $\langle U_n\rangle$ на $[0, 1)$, следующим образом определенная в двоичной записи:

$$\begin{aligned} U_0 &= 0.X_0 \\ U_1 &= 0.X_1X_2 \\ U_2 &= 0.X_3X_4X_5 \\ U_3 &= 0.X_6X_7X_8X_9 \\ &\ldots \end{aligned}$$

является случайной в смысле определения **R6**.
37. **[M48]** Существуют ли последовательности, удовлетворяющие определению **R4**, но не **R5**?
38. **[M50]** (**А. Н. Колмогоров.**) Пусть заданы N, n и ε. Найти наименьшее число алгоритмов в множестве A, таких, чтобы не существовали (n, ε)-случайные по отношению к A двоичные последовательности длины N. (Если нельзя получить точные формулы, можно ли найти асимптотические формулы? Главное в этой задаче—определить, насколько близка величина (35) к "наилучшей возможной".)
39. **[BM45]** (**К. Рот.**) Пусть $\langle U_n\rangle$—последовательность на $[0, 1)$ и $\nu_n(u)$—число неотрицательных целых чисел $j < n$, таких, что $0 \leq U_j < u$. Докажите, что существует положительная постоянная c, такая, что для любого N и любой последовательности $\langle U_n\rangle$ на $[0, 1)$ имеет место неравенство

$$|\nu_n(u) - un| > c\sqrt{\log N}$$

для некоторых n и u, таких, что $0 \leq n < N$, $0 \leq u < 1$. (Другими словами, никакая последовательность на $[0, 1)$ не может быть *слишком* равномерно распределенной.)

3.6. ВЫВОДЫ

В этой главе мы затронули большой круг вопросов: как вырабатывать случайные числа, как проверять их, как их преобразовывать и как получать связанные с ними теоретические результаты. Основной вопрос, который, наверное, возникнет по прочтении этой главы у большинства читателей, можно сформулировать следующим образом: "Каков же результат всей этой теории? Где простой и качественный датчик, который я мог бы использовать в своих программах как надежный источник случайных чисел?"

Из всего материала, изложенного в этой главе, следует, что "наилучший" и "простейший" датчик случайных чисел получается следующим образом. В начале программы установите переменную X равной некоторому целому значению X_0. Эта переменная X должна использоваться только для порождения случайных чисел. Как только для работы программы потребуется следующее случайное число, нужно установить

$$X \leftarrow (aX + c) \bmod m \tag{1}$$

и использовать новое значение X в качестве случайной величины. При выборе X_0, a, c и m необходимо соблюдать некоторые правила и использовать случайные числа осмысленно, руководствуясь следующими принципами.

i) Число X_0 может быть произвольным. Если по программе делается несколько просчетов и каждый раз желателен различный источник случайных чисел, присвойте X_0 последнее значение X, полученное во время предыдущего просчета, или (если это более удобно) установите X_0 равным текущему моменту времени и календарной дате.

ii) Число m должно быть велико. Удобно выбирать его равным размеру слова вычислительной машины, поскольку при этом эффективно вычисляется $(aX + c) \bmod m$. Вопросы, связанные с выбором m, более подробно разбираются в п. 3.2.1.1. Значение $(aX+c) \bmod m$ следует вычислять *точно*, без ошибок округления.

iii) Если m представляет собой степень двойки (т. е. если используется машина, работающая в двоичной системе счисления), выберите a таким, чтобы $a \bmod 8 = 5$. Если m есть степень 10 (т. е. используется машина, работающая в десятичной системе счисления), выберите a таким, чтобы $a \bmod 200 = 21$. При таком выборе величины a, при условии, что c выбирается описанным ниже способом, гарантируется, что датчик случайных чисел даст все m возможных различных значений X прежде, чем они начнут повторяться (см. п. 3.2.1.2), и, кроме того, гарантируется высокая "мощность" (см. п. 3.2.1.3).

iv) Множитель a должен превосходить величину $\sqrt{m}$, желательно, чтобы он был больше $m/100$, но меньше $m - \sqrt{m}$. Последовательность разрядов в двоичном или десятичном представлении a *не* должна иметь простого, регулярного вида. По этому поводу см., например, п. 3.2.1.3 и упр. 3.2.1.3-8. Лучше всего в качестве этого множителя выбрать наудачу некоторую постоянную, например,

$$a = 3141592621. \tag{2}$$

(Эта постоянная удовлетворяет обоим условиям из (iii).) Высказанных соображений обычно бывает достаточно для получения удовлетворительного множителя, однако, если датчик случайных чисел, используется интенсивно, множитель a, кроме того, следует выбирать так, чтобы удовлетворялся "Спектральный тест" (п. 3.3.4).

v) Постоянная c должна быть равна нечетному числу (когда m есть степень двойки) и кроме того не должна быть кратной 5 (когда m есть степень 10). Желательно выбирать c таким образом, чтобы отношение c/m было бы приблизительно равно величине, приведенной в п. 3.3.3, соотношение (41).

vi) Менее значимые (правые) разряды X не очень хороши с точки зрения случайности, поэтому при использовании числа X основную роль должны играть наиболее значимые разряды. Вообще говоря, лучше рассматривать X как случайную дробь X/m в интервале между 0 и 1, т. е. представлять X с десятичной точкой слева, чем как случайное целое число; расположенное между 0 и $m-1$. Для того, чтобы получить случайное целое число, расположенное между 0 и $k-1$, следует умножить X на k и отбросить лишние разряды (см. начало п. 3.4.2).

В этих шести "правилах" заключено все наиболее важное из того, что нужно знать о случайных числах, получаемых на вычислительной машине. К сожалению, в значительной части материала, опубликованного к моменту написания этой главы, рекомендуется использовать датчики, в которых наши правила нарушаются. Во многих статьях имеются важные теоретические результаты по выработке случайных чисел, однако предлагаемые в них конкретные методы несовершенны, что было обнаружено их авторами слишком поздно.

Возможно, что и рекомендуемые нами датчики случайных чисел в будущем будут признаны неудовлетворительными. Мы надеемся, что этого не произойдет, однако прошлое учит нас осторожности. Наиболее благоразумно было бы поступать так: прежде чем всерьез относиться к результатам расчета по программам, использующим метод Монте-Карло, следует считать по крайней мере дважды, используя совершенно различные источники случайных чисел. Такой подход не только гарантирует устойчивость результатов, но и контролирует качество датчика. (Любой датчик случайных чисел будет давать плохие результаты по крайней мере в одном классе приложений.)

Хотя линейные конгруэнтные последовательности, аналогичные описанным здесь, обычно являются удовлетворительным источником случайных чисел, в п. 3.3.4 (см. (3.3.4-13)) указано важное ограничение их качества. Из упр. 3.3.4-27 следует, что линейные конгруэнтные последовательности не годятся для расчетов по методу Монте-Карло с "высоким разрешением". Если требуется большая независимость случайных чисел, необходимо использовать методы п. 3.2.2.

Превосходный список литературы, опубликованный по этому вопросу до 1962 г., содержится в статье Т. Халда и А. Добелла "Random Number Generators" (*SIAM Review,* **4** (1962), 230–254). В связи с этим нам не нужно приводить ссылки на ранние публикации, кроме тех, которые уже упомянуты в этой главе.

Среди работ, появившихся после 1962 г., особенно следует отметить статью М. Макларена и Дж. Марсальи "Uniform random Generators" (*JACM,* **12** (1965), 83–89), потому что в ней впервые указаны недостатки линейных конгруэнтных датчиков с неправильно выбранными множителями. В

этой же статье был предложен алгоритм 3.2.2M. Это хороший метод, который значительно улучшает качество датчика и требует только вдвое большего времени счета.

Многие задачи, затронутые в настоящей главе, рассматриваются в книге Б. Янссона "Random Number Generators" (Stockholm, Almqvist and Wiksell, 1966). В книге Дж. Хаммерсли и Д. Хэндскомба "Monte Carlo Methods" (London, Methuen, 1964) подробно изучается применение случайных чисел в вычислительных методах. В этой книге показано, что эффективность многих численных методов возрастает, если использовать "квазислучайные" числа, специально подобранные для решения определенной задачи (и не обязательно удовлетворяющие статистическим тестам, о которых мы писали).

В приведенных ниже упражнениях имеются некоторые интересные проекты программ, использующих случайные числа. Возможно, что аналогичные проекты придут в голову читателю. По крайней мере один датчик случайных чисел имеется почти в любой хорошей программе.

Упражнения

1. [21] Напишите подпрограмму для машины MIX со следующими характеристиками:

Вызов:	`JMP RANDI`
Состояние при входе:	`rA` $= k$, положительное целое < 5000
Состояние при выходе:	`rA` $\leftarrow$ случайное целое Y, $1 \le Y \le k$, каждое целое значение равновероятно, `rX` $=?$; триггер переполнения сброшен.

(Используйте метод, предложенный в этом параграфе, но возьмите $c = 1$.)

>2. [21] Некоторые опасались, что вычислительные машины со временем будут управлять миром. Их успокаивают заявлением, что машина не может сделать ничего действительно нового, поскольку она всего лишь выполняет команды своего хозяина, программиста. Леди Лавлейс[13] писала в 1844 г.: "Аналитическая машина не притязает на *изобретение* чего-либо. Она может делать все то, *что мы сумеем ей приказать*". Многие философы развивали эту мысль. Обдумайте это утверждение, имея в виду датчики случайных чисел.

3. [32] "Игра в кости". Напишите программу, которая моделирует бросание двух костей, на каждой из которых с равной вероятностью выпадают значения $1, 2, \ldots, 6$. Если при первом бросании общая сумма равна 7 или 11, игра считается выигранной, если она равна 2, 3 или 12—проигранной. При любой другой сумме (назовем эту сумму "стрелкой") продолжайте бросания, пока не выпадет 7 (проигрыш) или опять получится "стрелка" (выигрыш).

Сыграйте десять игр. Результаты каждого бросания костей следует напечатать в виде mn, где m и n—цифры на двух костях, снабдив соответствующими комментариями вроде "змеиные глаза" и т. п.

4. [40] "Солитер" (пасьянс). Некоторые тратят драгоценное время на раскладывание пасьянса "солитер". Возможно, что автоматика вторгнется и в эту область. Напишите программу, которая (a) тасует смоделированную колоду карт, (b) раскладывает какой-либо обычный пасьянс "солитер", основываясь на полученном порядке карт в колоде, (c) печатает результат игры, откуда будет видно, насколько программа была близка к успеху. Следует сыграть несколько игр. Можно сделать так, чтобы по требованию программа начинала "мошенничать".

5. [46] *Моделирование литературного творчества с помощью вычислительной машины.* 26 октября 1960 г. по телевизионной программе CBS показывалась передача, озаглавленная "Думающая машина", в которой, в частности, использовались две небольшие пьесы в стиле вестерн, составленные при помощи программы для ЭВМ. Ниже приводятся два этих сценария в том виде, в котором их выдала машина.

---

[13] См. том 1, стр. 25.—*Прим. перев.*

Сценарий 1. (Револьвер в правой руке; деньги в левой руке; стакан на столе; бутылка на столе; кобура на грабителе; револьвер шерифа в правой руке шерифа; кобура шерифа на шерифе.)

ГРАБИТЕЛЬ (Грабитель за окном.) Идет к двери; открывает дверь; проходит в дверь; закрывает дверь; идет в угол; кладет деньги в углу; идет к столу; кладет револьвер на стол; садится и нервничает; сидит за столом; берет стакан (пустой) правой рукой; ставит стакан на стол; берет бутылку правой рукой; наливает; ставит бутылку на стол; берет стакан правой рукой; пьет из стакана; ставит стакан на стол; берет бутылку правой рукой; сидят за столом; сидит за столом; идет в угол; идет к окну; идет к столу; ставит бутылку на стол; садится и нервничает; сидит за столом; сидит и нервничает; идет к окну; идет к столу; берет стакан правой рукой.

ШЕРИФ (Шериф за окном.) Видит грабителя; (грабитель видит шерифа);

ГРАБИТЕЛЬ Ставит стакан на стол; берет револьвер правой рукой; проверяет револьвер.

ШЕРИФ Ждет; открывает дверь; видит грабителя; (грабитель видит шерифа); проходит в дверь.

ГРАБИТЕЛЬ Идет к окну; целится; стреляет; ЛЕГКО РАНИТ ШЕРИФА.

ШЕРИФ Идет к окну; целится; стреляет; ПРОМАХИВАЕТСЯ, идет к двери; идет к окну.

ГРАБИТЕЛЬ Идет к двери; целится; целится.

ШЕРИФ Целится; стреляет; ПРОМАХИВАЕТСЯ.

ГРАБИТЕЛЬ Стреляет; ЛЕГКО РАНИТ ШЕРИФА.

ШЕРИФ Идет к двери; целится; стреляет; ПРОМАХИВАЕТСЯ; проходит в дверь; целится.

ГРАБИТЕЛЬ Целится; стреляет; ПРОМАХИВАЕТСЯ; целится; стреляет; ПРОМАХИВАЕТСЯ.

ШЕРИФ Стреляет; ПРОМАХИВАЕТСЯ; идет к окну; целится; стреляет; ПРОМАХИВАЕТСЯ.

ГРАБИТЕЛЬ Целится; стреляет; ПРОМАХИВАЕТСЯ; целится; стреляет, ПРОМАХИВАЕТСЯ; целится; стреляет; ЛЕГКО РАНИТ ШЕРИФА.

ШЕРИФ Целится; стреляет; ТОЧНО ПОПАДАЕТ В ГРАБИТЕЛЯ.

ГРАБИТЕЛЬ Роняет револьвер; грабитель умирает.

ШЕРИФ Кладет револьвер в кобуру; идет к столу; берет стакан (пустой) правой рукой; берет левой рукой стакан из правой руки; берет бутылку правой рукой; наливает; ставит бутылку на стол; берет правой рукой стакан из левой руки; пьет из стакана; берет левой рукой стакан из правой руки; берет бутылку правой рукой; наливает; ставит бутылку на стол; берет правой рукой стакан из левой руки; пьет из стакана; ставит стакан на стол; идет в угол; берет деньги правой рукой; идет к двери, проходит в дверь; закрывает дверь. ЗАНАВЕС.

Сценарий 2. (Револьвер в правой руке; деньги в левой руке; стакан на столе; бутылка на столе; кобура на грабителе; револьвер шерифа в правой руке шерифа; кобура шерифа на шерифе.)

ГРАБИТЕЛЬ (Грабитель за окном.) Идет к двери; открывает дверь; проходит в дверь; закрывает дверь; идет в угол; кладет деньги в углу; идет к окну; кладет револьвер у окна; облокачивается и смотрит в окно; облокачивается и смотрит в окно, идет в угол; считает деньги; идет к столу; берет стакан (пустой) правой рукой; берет левой рукой стакан из правой руки; берет бутылку правой рукой; наливает; ставит бутылку на стол; берет правой рукой стакан из левой руки; пьет из стакана; ставит стакан на стол; берет бутылку правой рукой; наливает; идет в угол; ставит бутылку в углу; идет к окну; берет револьвер правой рукой; проверяет револьвер; кладет револьвер в кобуру; идет к столу; берет стакан правой рукой; пьет из стакана; идет к окну; ставит стакан у окна.

ШЕРИФ (Шериф за окном.) Видит грабителя; (грабитель видит шерифа); идет к двери.

ГРАБИТЕЛЬ Берет револьвер из кобуры правой рукой; проверяет револьвер, идет к двери; проверяет револьвер; кладет револьвер у двери.

ШЕРИФ Открывает дверь; видит грабителя; (грабитель видит шерифа); проходит в дверь; идет к окну.

ГРАБИТЕЛЬ Берет револьвер правой рукой.

ШЕРИФ Идет к столу.

ГРАБИТЕЛЬ Целится; стреляет; ПРОМАХИВАЕТСЯ; целится; стреляет; ТОЧНО ПОПАДАЕТ В ШЕРИФА; продувает ствол; кладет револьвер в кобуру.

ШЕРИФ Роняет револьвер; шериф умирает.

ГРАБИТЕЛЬ Идет в угол; берет деньги правой рукой; идет к двери; проходит в дверь; закрывает дверь. ЗАНАВЕС.

Если внимательно прочитать сценарии, можно заметить, что они полны напряженного драматизма. Программа достаточно точно определяла, где находится каждое действующее лицо, что оно держит в руках и т. д. Действия актеров случайны и подчиняются определенным вероятностям, вероятность неразумных действий возрастает в зависимости от того, сколько выпито виски и сколько

раз действующее лицо было ранено. Читатель может изучить приведенные сценарии для того, чтобы выявить другие свойства программы.

Разумеется, даже лучшие сценарии приходится переделывать, прежде чем поставить, и тем более это необходимо сделать со сценарием, написанным неопытным автором. Ниже приведены сценарии в том виде, в котором их использовали в телевизионной передаче (КП—крупный план, СП-средний план, ДП—дальний план).

Сценарий 1. Музыка.

СП. Грабитель смотрит в окно хижины.
КП. Лицо грабителя.
СП. Грабитель входит в хижину.
КП. Грабитель видит на столе бутылку виски.
КП. Шериф снаружи хижины.
СП. Грабитель видит шерифа.
ДП. Шериф в дверном проеме виден над плечом грабителя, оба хватаются за кобуры.
СП. Шериф выхватывает револьвер.
ДП. Стреляет, попадает в грабителя.
СП. Шериф берет мешки с деньгами.
СП. Грабитель шатается.
СП. Грабитель умирает. Пытается последний раз выстрелить в шерифа, падает на стол.
СП. Шериф выходит из двери с деньгами.
СП. Тело грабителя, неподвижно лежащее на столе. Камера двигается назад. (Смех.)

Сценарий 2. Музыка.

КП. Окно. Появляется грабитель.
СП. Грабитель входит в хижину с двумя мешками денег.
СП. Грабитель кладет мешки с деньгами на бочку.
КП. Грабитель видит на столе виски.
СП. Грабитель наливает себе виски. Начинает считать деньги. Смеется.
СП. Шериф снаружи хижины.
СП. Вид через окно.
СП. Грабитель видит шерифа через окно.
ДП. Шериф входит в хижину. Хватается за револьвер. Выстрел.
КП. Шериф. Корчится от боли.
СП. Шериф, шатаясь идет к столу, чтобы выпить, ... падает мертвый.
СП. Грабитель уходит из хижины с мешками денег.

Вышеизложенное любезно сообщили автору Т. Вулф, постановщик телевизионной программы, предложивший идею написания небольшой пьесы с помощью ЭВМ, а также Д. Росс и Х. Морс, которые написали и отладили программу. [14]

Без сомнения, у читателя возникнут свои соображения о том, как бы он мог написать программу моделирования литературного творчества. Это и есть цель данного упражнения.

4. Арифметика

> Ввиду того что нет ничего более хлопотного в математической практике (о возлюбленные студенты-математики!), ничего такого, что более досаждало бы и мешало вычислителю, чем выполняемые над большими числами умножение, деление, извлечение квадратных и кубических корней, которые сопряжены обычно с массой трудно обнаруживаемых ошибок, я стал размышлять над тем, какое надежное и удобное средство могло бы устранить эти помехи.
> Джон Непер (1614)

[14] Для дальнейших справок см. J. E. Pfeiffer, The Thinking Machine (New York: J. B. Lippincott, 1962).

Терпеть не могу складывать. Нет большей ошибки, чем называть арифметику точной наукой. Существуют ... тайные законы, управляющие числами, постичь которые может лишь ум типа моего. Например, если вы находите сумму, складывая числа столбиком сначала снизу вверх, а затем сверху вниз, вы всегда получаете разный результат.
Госпожа Ла Туш (19 в.)

Не могу представить себе, чтобы кому-нибудь потребовалось выполнять умножение со скоростью $40\ 000$ или даже $4\ 000$ операций в час; такое радикальное изменение, как переход к восьмеричной системе, не следует навязывать всему человечеству ради нескольких личностей.
Ф. Х. Уэйлс (1936)

Основная цель этой главы—тщательное рассмотрение четырех основных действий арифметики: сложения, вычитания, умножения и деления. Многие считают арифметику тривиальной вещью, которой обучают детей в школе и действия которой выполняют машины, но, как мы увидим, арифметика — это увлекательный предмет, имеющий много интересных аспектов. Арифметика лежит в основе столь большого числа машинных приложений, что важно самым доскональным образом исследовать эффективные методы вычислений с числами.

Арифметика—это на самом деле живая и все еще быстро развивающаяся область науки, сыгравшая важную роль в мировой истории. В этой главе мы проанализируем алгоритмы выполнения арифметических операций над многими типами величин, такими, как числа "с плавающей точкой", очень большие числа, дроби (рациональные числа), многочлены и степенные ряды; мы обсудим также связанные с этим вопросы, такие, как преобразование из одной системы счисления в другую, разложение чисел на множители и вычисление многочленов.

4.1. ПОЗИЦИОННЫЕ СИСТЕМЫ СЧИСЛЕНИЯ

То, каким образом мы выполняем арифметические действия, тесно связано с тем, каким образом мы представляем числа, с которыми работаем; поэтому наше изучение арифметики естественно начать с обсуждения принципиальных способов представления чисел.

Позиционное представление с *основанием* (или *по основанию*) b определяется правилом

$$(\dots a_3a_2a_1a_0.a_{-1}a_{-2}\dots)_b = \dots + a_3b^3 + a_2b^2 + a_1b^1 + a_0 + a_{-1}b^{-1} + a_{-2}b^{-2} + \dots; \qquad (1)$$

например, $(530.3)_6 = 5\cdot 6^2 + 2\cdot 6^1 + 0 + 3\cdot 6^{-1} = 192\frac{1}{2}$. Наша традиционная десятичная система счисления—это, разумеется, тот частный случай, когда b равно десяти и когда значения a выбираются из "десятичных цифр" $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$; в этом случае индекс b в (1) можно опускать.

Простейшее обобщение десятичной системы счисления получается, когда в качестве b берут любое целое число, большее единицы, и числа a—это целые числа из интервала $0 \le a_k < b$. Так получаются стандартные двоичная ($b = 2$), троичная ($b = 3$), четверичная ($b = 4$), пятеричная ($b = 5$), ... системы счисления. Более общо, в качестве b можно было бы взять произвольное число, а числа a выбирать из произвольного заранее заданного множества чисел; как мы увидим, это приводит к некоторым интересным ситуациям.

Точка, стоящая между a_0 и a_{-1} в (1), называется *позиционной* (или *разделительной*) точкой. (В случае $b = 10$ ее называют также десятичной точкой, в случае $b = 2$—двоичной точкой и т. д.) В европейских странах[1] вместо позиционной точки часто используют запятую[2].

Числа a в (1) называют *цифрами* представления. Цифру a_k с бóльшим k называют "более значимой", чем цифру a_k с меньшим k; соответственно самую левую (*ведущую*, или *головную*) цифру называют *наиболее значимой цифрой*, а самую правую (*хвостовую*)—*наименее значимой*. В стандартной двоичной системе двоичные цифры часто называют *битами*. В стандартной шестнадцатеричной

[1] Кроме Англии.—*Прим. перев.*

[2] Точка вместо запятой постепенно входит в употребление и у нас в связи с распространением алгоритмических языков программирования "англосакского происхождения" типа АЛГОЛ, ФОРТРАН или ПЛ/1.—*Прим. перев.*

системе (с основанием шестнадцать) шестнадцатеричные цифры от нуля до пятнадцати обозначают обычно так:

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, \mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}.$$

История развития способов представления чисел—это увлекательнейшая повесть, действие которой происходит параллельно с развитием самой цивилизации. Рассматривая, однако, эту историю во всех подробностях, мы недопустимо далеко отошли бы от нашей главной темы; тем не менее будет полезно дать набросок основных ее моментов.

Наиболее ранние способы представлений чисел, которые до сих пор можно наблюдать в примитивных цивилизациях, основаны на использовании групп пальцев или кучек камней и т. п., обычно с дополнительными соглашениями относительно замены некоторой большой кучки или группы, скажем из пяти или десяти объектов, одним объектом специального вида или расположенным в специальном месте. Подобные системы естественно приводят к наиболее ранним способам представления чисел в письменной форме, таким, как египетские, вавилонские, греческие[3], китайские и римские числа, но эти обозначения чрезвычайно неудобны для выполнения арифметических действий, за исключением простейших случаев.

Проводившиеся историками математики в двадцатом столетии широкие исследования древних клинописных табличек, найденных археологами на Ближнем Востоке, показали, что вавилоняне применяли в действительности две различные системы представления чисел. Числа, использовавшиеся для повседневных деловых заметок, записывались при помощи обозначений основанных на группировании по десяткам, сотням и т. д., унаследованных от более ранних цивилизаций Месопотамии; большие числа требовались здесь редко. Когда же вавилонским математикам приходилось рассматривать более трудные математические задачи, они широко применяли шестидесятеричную позиционную систему счисления (по основанию шестьдесят), которая была хорошо развита уже ко времени не позднее 1750 г. до н. э. Эта числовая система была уникальной в том отношении, что она была фактически формой представления с плавающей точкой с опущенным показателем; нужный масштабный множитель, т. е. степень шестидесяти, следовало определять из контекста так что например, все числа из списка 2, 120, 7200, $1/30$ и т. д. записывались одинаковым образом. Эта система была в особенности удобна для выполнения умножения и деления при помощи вспомогательных таблиц, так как выравнивание порядков никак не влияло на ответ; та же идея реализована ныне в логарифмической линейке. Примерами этой вавилонской системы записи могут служить следующие извлечения из древних таблиц: квадрат 30 есть 15 (что можно прочитать также как "квадрат $1/2$ равен $1/4$"); число, обратное к $81 = (1\ 21)_{60}$, равно $(44\ 26\ 40)_{60}$; квадрат этого последнего числа равен $(32\ 55\ 18\ 31\ 6\ 40)_{60}$. У вавилонян был особый символ для нуля, но из-за их идеологии плавающей точки нуль использовался только внутри чисел и никогда в крайней правой позиции для обозначения масштаба. Об интересной истории развития ранней вавилонской математики можно прочитать у О. Нойгебауэра [The exact sciences in antiquity, Princeton University Press, 1952] и Б. Л. Ван дер Вардена [Пробуждающаяся наука, Физматгиз, 1959].

Позиционное представление чисел с фиксированной точкой, во-виднмому, впервые было развито индейцами племени майя в Центральной Америке около 2000 лет тому назад; их система счисления с основанием 20 была достаточно высокоразвита, особенно в отношении записи астрономических фактов и календарных дат. Однако испанские завоеватели уничтожили почти все исторические и научные тексты майя, поэтому мы не знаем, насколько далеко продвинулись они в арифметике; найдены некоторые таблицы умножения специального назначения, но неизвестно никаких примеров деления [см. J. Eric S. Thompson, *Contributions to Amer. Antropology and History,* Carnegie Inst. of Washington, **7** (1942), 37–62].

За несколько веков до новой эры греки применяли для своих арифметических вычислений ранюю разновидность абаки, в которой использовались песок и/или галька на доске, имевшей ряды или столбцы, естественным образом соответствовавшие нашей десятичной системе. Нам, возможно, покажется удивительным, что тот же самый позиционный принцип не был применен ими для записи чисел—ведь мы так привыкли проводить расчеты в десятичной системе при помощи карандаша и бумаги; но большая легкость вычислений на абаке (писать тогда умели не все, и, кроме того, вычисления на абаке делали ненужным запоминание таблиц сложения и умножения) привела, вероятно, греков к убеждению, что нелепо даже предполагать, что вычисления можно удобнее проводить, "царапая на бумаге". В то же время греческие астрономы использовали для записи дробей шестидесятеричную позиционную систему счисления, чему они научились у вавилонян.

---

[3] К греческой числовой системе восходят и старославянские числовые обозначения, использовавшиеся главным образом в хронологических целях.—*Прим. перев.*

Наша десятичная система, отличающаяся от более древних форм в первую очередь наличием позиционной точки вместе с символом нуль для обозначения пустой позиции, впервые появилась у индусов. Точная дата возникновения этой системы неизвестна; есть основания предполагать, что это произошло примерно в 6 в. н. э. Индийская наука того времени, особенно астрономия, отличалась весьма высоким уровнем. В наиболее ранних из дошедших до нас индийских рукописей, использовавших эту систему счисления, числа записываются в обратном порядке (с наиболее значимой цифрой справа), но вскоре стало правилом располагать наиболее значимую цифру слева.

Около 750 г. н. э. индусские принципы десятичной арифметики распространились в Персии, когда на арабский язык было переведено несколько важных работ; яркая картина этого периода дана в одной древнееврейской рукописи, перевод которой на английский опубликован в журнале *AMM* [**15**, (1918), 99–108]. Вскоре после этого аль-Хорезми написал на арабском свое руководство по этому предмету. (Как отмечалось в гл. 1, от его имени произошло наше слово "алгоритм".) Книга аль-Хорезми была переведена на латынь и оказала сильное влияние на Леонардо Пизанского (Фибоначчи), чья книга по арифметике (1202 г.) в свою очередь сыграла решающую роль в распространении индо-арабских чисел в Европе. Интересно отметить, что в результате этих двух переводов—с санскрита на арабский и с арабского на латынь—порядок написания чисел (слева направо) не изменился, хотя арабы пишут справа налево, а индусы и европейцы слева направо. Подробное описание последующего процесса распространения десятичных обозначений и арифметики по всей Европе в период с 1200 по 1600 г. дано в книге [David Eugene Smith, History of mathematics, **1**, Boston, Gihn and Co., 1923], гл. 6 и 8.

Десятичная система счисления применялась вначале только к целым числам, а для дробей не использовалась. Арабские астрономы, которым приходилось работать с дробями при составлении карт звездного неба и других таблиц, продолжали пользоваться обозначениями Птолемея (знаменитого греческого астронома), основанными на шестидесятеричных дробях. Эта система дожила до наших дней в наших единицах углов ("градусы, минуты и секунды"), а также в единицах времени—рудиментах первоначальной шестидесятеричной системы вавилонян. Первые европейские математики, когда им приходилось иметь дело с нецелыми числами, также пользовались шестидесятеричными дробями; Фибоначчи, например, приводил в качестве приближенного корня уравнения $x^3 + 2x^2 + 10x = 20$ значение

$$1^\circ\, 22'\, 7''\, 42'''\, 33^{IV}\, 4^{V}\, 40^{VI}.$$

Использование десятичных обозначений для десятых, сотых и т. д. кажется сравнительно небольшим продвижением, но, конечно, разрыв с традицией всегда труден, а кроме того, шестидесятеричные дроби имеют преимущество перед десятичными в том, что такие числа, как $1/3$, допускают точную запись в простом виде.

Первыми, кто стал работать с величинами, эквивалентными десятичным дробям, были китайские математики (которые никогда не пользовались шестидесятеричной системой), хотя их числовая система (без нуля) и не была первоначально позиционной системой в строгом смысле слова. Китайские единицы весов и мер были десятичными, так что Цзу Чун-чжи (умер около 500 г.) смог выразить приближение числа π как "3 чана, 1 чжи, 5 цуня, 1 фэн, 5 ли, 9 хао, 2 мяо, 7 ху". Здесь чан, ..., ху—единицы длины; 1 ху (диаметр шелковой нити) равен $1/10$ мяо и т. д. Использование таких похожих на десятичные дробей получило довольно широкое распространение в Китае начиная примерно с 1250 г. В истинно позиционной системе десятичные дроби впервые появляются в трактате по арифметике, написанном в Дамаске неизвестным математиком, подписавшимся "аль-Уклидиси" ("Евклидов"). Он ввел символ $'$ для обозначения десятичной точки и указал выражения для некоторых дробей и приближенных квадратных и кубических корней; но он не понял, что целую и дробную части чисел можно умножать одновременно. Его работа не стала общеизвестной, и через пять веков десятичные дроби были заново открыты персидским математиком аль-Каши, умершим около 1436 г. Аль-Каши был высоко искусным вычислителем, который дал следующее значение для 2π, содержащее 16 правильных, десятичных знаков:

Целая часть		Дробная часть														
0	6	2	8	3	1	8	5	3	0	7	1	7	9	5	8	6	5

Это было наилучшим приближением для числа π до тех пор, пока Лудольф ван Цейлен после многих трудов в период с 1596 по 1610 г. не вычислил 35 десятичных знаков. Самые ранние из известных примеров десятичных дробей в Европе обнаружены в одном тексте 15 в., где, например, 153.5 умножается на 16.25 и в ответе получается 2494.375; это названо там "турецким методом". В 1525 г. Кристоф Рудольф из Германии самостоятельно открыл десятичные дроби; его работа также осталась неизвестной. Франсуа Вьет предлагает ту же идею снова в 1579 г. Наконец, трактат по арифметике, написанный Симоном Стевином из Нидерландов, в свою очередь независимо в 1585 г. пришедшим к идее десятичных дробей, приобретает широкую популярность.

Его работа и последовавшее вскоре открытие логарифмов привели к тому, что в 17 в. десятичные дроби стали общепринятыми в Европе. [Дальнейшие подробности и ссылки на литературу можно найти в книгах: D. E. Smith, History of Mathematics, **2**, Boston, Ginn and Co., 1925, 228–247, и C. B. Boyer, History of Mathematics, New York, Wiley, 1968[4].]

Двоичная система счисления имеет свою собственную интересную историю. Известно, что многие примитивные племена, существующие в наше время, используют двоичную, или "парную", систему счета (группировка по парам, а не по пятеркам или десяткам), но было бы неверным сказать, что они вычисляют в настоящей двоичной системе счисления, так как они не выделяют специальным образом степеней двойки. Интересные подробности о примитивных вычислительных системах можно найти в статье Абрахама Сайденберга "The diffusion of counting practices" [*Univ. Calif. Publ. in Math.*, **3** (1960), 215–300]. Другой "примитивный" пример существенно двоичной системы—это традиционные обозначения для длительности нот в музыке.

В папирусе Ринда[5](Египет, около 1650 г. до н. э.), одном из первых известных нетривиальных математических документов, применяется десятично ориентированная схема представления чисел, но в нем же показано, как выполнять операцию умножения с помощью последовательных удвоений и сложений. Эта процедура по существу своему основана на двоичном представлении чисел, хотя двоичная система счисления специально и не вводилась.

В 17 в. недесятичные счисления становятся предметом рассмотрения в Европе. В течение многих лет астрономы от случая к случаю использовали шестидесятеричную арифметику как для целых чисел, так и для дробных частей, главным образом при выполнении умножения; см. книгу Джона Валлиса "Treatise of algebra" [Oxford, 1685], 18–22, 30. Тот факт, что *любое* положительное число может служить основанием системы счисления, был, по-видимому, впервые опубликован в работе Блеза Паскаля "De numeris multiplicibus", написанной около 1658 г. [см. B. Pascal, Euvres Complètes, Paris, Éditions de Seuil, 1963, 84–89]. Паскаль писал: "Denaria enim ex institute hominum, non ex necessitate naturae ut vulgus arbitratur, et sane satis inepte, posita est", т. е.: "Десятичная система построена—довольно неразумно, конечно,—в соответствии с людскими обычаями, а вовсе не требованиями естественной необходимости, как склонно думать большинство людей". Он утверждал, что было бы желательно перейти к двенадцатеричной системе, и предложил правило проверки делимости двенадцатеричного числа на 9. Четверичную систему счисления в ряде публикаций начиная с 1673 г. предлагал Эрхард Вайгель. Подробное обсуждение арифметики по основанию двенадцать было проведено Джошуа Джордэйном [Duodecimal arithmetick, London, 1687].

Хотя на протяжении всей этой эпохи в арифметике применялась почти исключительно десятичная система, системы мер и весов редко когда, если вообще когда-либо, строились на основе кратных десяти, и многие деловые операции требовали изрядной ловкости в сложении величин типа фунтов, шиллингов и пенсов. Итак, столетиями купцы учились вычислять суммы и разности величин, выраженных в специфических денежных единицах или единицах мер и весов, а это фактически была арифметика в недесятичной системе счисления. Особого внимания заслуживают, в частности, распространенные в Англии единицы измерения жидкостей, установившиеся еще в 13 в. или даже раньше:

2 джилла = 1 полуштоф,
2 полуштофа = 1 пинта,
2 пинты = 1 кварта,
2 кварты = 1 потл,
2 потла = 1 галлон,
2 галлона = 1 пек,
2 пека = 1 полубушель,
2 полубушеля = 1 бушель, или фиркин,
2 фиркина = 1 килдеркин,
2 килдеркина = 1 баррель,
2 барреля = 1 хогзхед,
2 хогзхеда = 1 пайп,
2 пайпа = 1 тан.

---

[4] См. также книгу Д. Я. Стройка "Краткий очерк истории математики", "Наука", М., 1950.— *Прим. перев.*

[5] Назван по имени его владельца—английского египтолога Ринда (A. H. Rhind). Хорошей репродукцией фрагмента этого папируса открывается сборник "Математика в современном мире" ("Мир", М., 1967).— *Прим. ред.*

Объемы жидкости, выраженные в галлонах, потлах, квартах, пинтах и т. д.[6], записывались по существу в двоичной системе. Быть может, подлинными изобретателями двоичной арифметики были английские виноторговцы!

Насколько сейчас известно, впервые двоичная система счисления появляется около 1605 г. в ряде неопубликованных рукописей Томаса Хэрриота (1560–1621). Хэрриот—весьма творческая личность—прибыл в Америку в качестве представителя сэра Уолтера Рэйли. Он изобрел (среди многого другого) используемые ныне обозначения для отношений "меньше" и "больше"; но по некоторым соображениям он предпочел не публиковать многие из своих открытий. Извлечения из его заметок по двоичной арифметике были воспроизведены Джоном У. Шэрли в *Amer. J. Physics* [**19** (1951), 452–454].

Первое опубликованное обсуждение двоичной системы счисления принадлежит испанскому священнику Хуану Карамюэлю Лобковицу; это сравнительно малоизвестная работа [*Mathesis Bicepts,* Campaniæ, **1** (1670), 45–48]. Карамюэль довольно подробно рассмотрел представление чисел в системах по основаниям 2, 3, 4, 5, 6, 7, 8, 9, 10, 12 и 60, но не привел никаких примеров арифметических операций в недесятичных системах (за исключением тривиальной операции прибавления единицы).

Наконец, одна статья Г. В. Лейбница [*Memoires de l'Academic Royale des Sciences,* Paris, (1703), 110–116], в которой пояснялись двоичные операции сложения, вычитания, умножения и деления, действительно привлекла к двоичной системе всеобщее внимание, и именно на эту статью обычно ссылаются, говоря о рождении арифметики по основанию 2. Лейбниц и далее очень часто обращался к двоичной системе счисления [см. W. Ahrens, Mathematische Unterhaltungen und Spiele, **1**, Leipzig, Teubner, 1910, 27–28]. Он не рекомендовал ее для практических вычислений, но подчеркивал ее важность в теоретико-числовых исследованиях, так как закономерности поведения числовых последовательностей часто гораздо легче усмотреть в двоичной записи, нежели в десятичной; он видел также некий мистический смысл в том факте, что всё на свете выразимо с помощью нулей и единиц. [См. Laplace, Théorie analytique des Probabilités, 3^{me} éd., Paris, 1820, cix].

Интересно отметить, что важная концепция отрицательных степеней справа от позиционной точки не была еще в те времена по-настоящему осознана. Лейбниц просил Якова Бернулли вычислить π в двоичной системе счисления, и Бернулли "решил" задачу, взяв 35-значное приближение к π, умножив его на 10^{35}, а затем выразив полученное целое число в двоичной системе; это и был его ответ! Для меньшего масштабного множителя это рассуждение выглядело бы так: $\pi \approx 3.14$, а $(314)_{10} = (100111010)_2$, следовательно, π в двоичной системе счисления есть 100111010![7]. (См. Leibniz, Math. Schriften (ed. Gehrhardt), **3** [Halle, 1855], 97; из-за ошибок в вычислениях два из 118 битов в ответе неверны.) Цель вычислений Бернулли состояла, по-видимому, в том, чтобы выяснить, можно ли обнаружить в этом представлении π какие-нибудь простые закономерности.

Шведский король Карл XII, математический талант которого, если его сравнивать с математическими талантами прочих королей, по-видимому, не имел себе равных во всей мировой истории, около 1717 г. увлекся идеей восьмеричной арифметики. Скорей всего это было его собственное изобретение, хотя он и встречался с Лейбницем в 1707 г. Карл чувствовал, что основание 8 или 64 было бы более удобным для вычислений, чем 10, и подумывал о введении в Швеции восьмеричной арифметики; однако он погиб в битве, так и не успев провести эту реформу. (См. сочинения Вольтера, **21** [Paris, E. R. DuMont, 1901], 49; E. Swedenborg, *Gentleman's Magazine,* **24** (1754), 423–424.)

Примерно 140 лет спустя выдающийся американский гражданский инженер, швед по национальности Джон Нистром решил сделать еще один шаг в развитие планов Карла XII и предложил полную систему нумерации, мер и весов, основанную на шестнадцатеричной арифметике. Он писал: "Я не боюсь и ничуть не колеблюсь выступить в защиту двоичной системы в арифметике и метрологии. Я знаю, на моей стороне природа; если мне не удастся убедить вас в ее полезности и чрезвычайной важности для человечества, это не сделает чести нашему поколению, нашим ученым и философам". Нистром разработал специальные правила произнесения шестнадцатеричных чисел; например, число $(\texttt{C0160})_{16}$ следовало читать "vybong, bysanton". Полностью эта система, названная им "тональной системой", была описана в J. Franklin Inst. [**46** (1863), 263–275, 337–348, 402–407]. Аналогичная система, но использующая основание 8, была предложена примерно в то же время Элфредом Б. Тэйлором [*Proc. Amer. Pharmaceutical Assoc.,* **8** (1859), 115–216; *Proc. Amer. Philosophical Soc.,* **24** (1887), 296–366].

Все более широкое применение французской (метрической) системы мер и весов вызвало в ту эпоху многочисленные дебаты о достоинствах десятичной арифметики.

---

[6] Происхождение названий этих мер объема довольно прозрачно: firkin, например, означает "маленький бочонок"; tan—"большая бочка".—*Прим. перев.*

[7] Это запись не π, а $10^2\pi$! Вот первые цифры числа π в двоичной системе: $11.001001000011111011\ldots$.—*Прим. ред.*

Со времени Лейбница двоичная система счисления становится хорошо известной диковинкой, и Р. К. Арчибалд собрал около 20 ранних работ, посвященных ей [*AMM,* **25** (1918), 139–142]. Она применялась главным образом при вычислении степеней, как будет объяснено в п. 4.6.3, и при анализе некоторых игр и головоломок. Дж. Пеано [*Atti della R. Accademia delle Scienze di Torino,* **34** (1898), 47–55] использовал двоичную систему как основу для "логического" алфавита из 256 символов. Джозеф Боуден [Special topics in theoretical arithmetic. Garden City, 1936, 49] предложил систему обозначений для шестнадцатеричных чисел.

Возрастающий интерес к механическим устройствам для выполнения арифметических операций, в особенности умножений, привел ряд исследователей к изучению двоичной системы с этой точки зрения. Прекрасный отчет об этих исследованиях дан в статье "Двоичные вычисления" Э. У. Филлипса [*J. of the Institute of Actuaries,* **67** (1936), 187–221]; там же помещена запись дискуссии, состоявшейся после прочитанной им на эту тему лекции. Филлипс начинает словами: "Конечная цель [этой статьи] состоит в том, чтобы убедить весь цивилизованный мир отказаться от десятичной нумерации и заменить ее восьмеричной[8]".

Современные читатели статьи Филлипса будут, возможно, удивлены, обнаружив, что система счисления по основанию 8 называлась в то время в соответствии со всеми словарями английского языка octonary или octonal, точно так же как система по основанию 10 называется сейчас denary или decimal; слово octal появляется в словарях английского языка только с 1961 г., причем первоначально, по-видимому, как термин при описании базы вакуумных электронных ламп определенного класса. Слово hexadecimal[9], вкравшееся в наш язык еще позже, представляет собой смесь греческого и латинского корней[10]; более последовательными терминами были бы senidenary[11] или sedecimal или даже sexadecimal[12], но последний, пожалуй, звучал бы слишком рискованно для программистов.

Высказывание Ф. Х. Уэйлса, приведенное в качестве одного из эпиграфов к этой главе, извлечено как раз из записи дискуссии, помещенной вместе со статьей Филлипса. Другой слушатель этой лекции указал на неудобства восьмеричной системы для деловых целей: "5% превращаются в 3.1463 per 64[13], что звучит ужасно".

Ряд вычислительных машин, основанных на двоичной системе был создан в начале тридцатых годов нашего столетия во Франции; см. заметки Л. Куффиньяля и Р. Вальта в *Comptes Rendus* [**197** (1933), 877; **202** (1936), 1745–1747, 1970–1972].

Первые вычислительные схемы на электронных лампах спроектировал в 1937 г. Джон В. Атанасов, а первые релейные вычислительные схемы независимо и в том же году—Джордж Р. Штибиц. Оба они в своих проектах использовали двоичную систему счисления, хотя Штибиц вскоре после этого разработал и двоичный код "плюс-3" для десятичных цифр. Примерно в то же самое время в Германии Конрад Цузе построил механическую вычислительную машину, основанную на двоичном представлении чисел с плавающей точкой; впоследствии (1941 г.) он заменил механическую логическую схему релейной схемой, которая заработала в 1941 г.

В первых быстродействующих вычислительных машинах, построенных в Америке в начале сороковых годов, использовалась десятичная арифметика. Но в 1946 г. в сыгравшем большую роль отчете А. У. Бёркса, Х. Х. Голдстайна и Дж. фон Неймана о проекте первой вычислительной машины с хранимой в памяти программой были подробно изложены причины их решения порвать с традицией и перейти к системе счисления по основанию 2 [см. John von Neumann, Collected Works, Vol. 5, 41–65]. С тех пор двоичные вычислительные устройства получили всеобщее распространение. После первой дюжины лет работы с двоичными машинами обсуждение сравнительных достоинств и недостатков двоичной системы было дано В. Буххольцем в статье "Пальцы или кулаки?"[14] [*CACM,* **2** (December, 1959), 3–11].

Вычислительная машина MIX, используемая в этой книге, определена таким образом, что она может быть как двоичной, так и десятичной. Интересно отметить, что почти все MIX-программы можно записать, не зная, какая именно система используется, двоичная или десятичная,—даже при проведении вычислений многократной точности. Итак, мы видим, что выбор основания системы

---

[8] В оригинале octonal, как у самого Филлипса, а не octal, как принято в современном английском языке. В следующем абзаце речь идет о читателях английского оригинала.—*Прим. ред.*

[9] Шестнадцатеричный.—*Прим. перев.*

[10] Вот эти корни: $\tilde{\varepsilon}\xi$—шесть, decimus—десятый. *Прим. ред.*

[11] Seni deni—шестнадцать (*лат.*).—*Прим ред.*

[12] Sex—шесть (*лат.*).—*Прим. ред.*

[13] Поскольку % пишется по-английски per cent (от латинского per centum), а cent (100) заменяется на 64.—*Прим. ред.*

[14] Автор статьи обыгрывает версию антропологического происхождения десятичной системы, подыскивая антропологическое обоснование и для двоичной системы.—*Прим. перев.*

счисления не оказывает серьезного влияния на программирование для ЭВМ. (Заслуживающим упоминания исключением из этого правила служат, однако, "булевы" алгоритмы, обсуждаемые в гл. 7; см. также алгоритм 4.5.2B.)

Имеется несколько различных методов представления *отрицательных* чисел в ЭВМ, и выбор того или иного метода оказывает влияние на способы реализации арифметических действий. Разберем различие между этими обозначениями. Будем сначала считать машину MIX десятичной ЭВМ; тогда каждое слово содержит 10 цифр и знак, например:

$$-12345\,67890. \tag{2}$$

Этот способ представления называется *прямым кодом*. Такое представление соответствует общепринятым обозначениям, и поэтому многие программисты предпочитают его. Возможное неудобство здесь состоит в том, что могут появляться как минус нуль, так и плюс нуль, в то время как обычно они должны обозначать одно и то же число; такая возможность требует принятия некоторых мер предосторожности.

В большинстве механических счетных машин, выполняющих действия десятичной арифметики, используется другая система записи, называемая *дополнительным кодом*. Вычтя 1 из $00000\,00000$, мы получим в этой системе записи $99999\,99999$; другими словами, числу не приписывается явного знака, а вычисления проводятся *по модулю* 10^{10}. Число $-12345\,67890$ в дополнительном коде будет выглядеть так:

$$87654\,32110. \tag{3}$$

В этой системе обозначений принято считать отрицательным любое число, головная цифра которого есть 5, 6, 7, 8 или 9, хотя с точки зрения правильности результатов сложения и вычитания не будет никакого греха рассматривать (3), если это удобно, как число $+87654\,32110$. При применении дополнительного кода не возникает и проблемы "минус нуля". Главное различие между прямым кодом и дополнительным состоит практически в том, что сдвиг вправо в дополнительном коде не эквивалентен делению на 10; например, число $-11 = \ldots 99989$ после сдвига вправо превращается в число $\ldots 99998 = -2$ (в предположении, что сдвиг вправо отрицательного числа порождает в головном разряде "9"). В общем случае результатом сдвига числа x, записанного в дополнительном коде, на одну цифру вправо будет число $\lfloor x/10 \rfloor$ независимо от того, положительно x или отрицательно. Одно из потенциальных неудобств записи при помощи дополнительного кода заключается в ее несимметричности относительно нуля; наибольшее отрицательное число, представимое посредством p цифр, $500\ldots0$, не является знаковым обращением никакого p-разрядного положительного числа. Таким образом, возможно, что изменение знака (замена x на $-x$) послужит причиной переполнения.

Еще одна система обозначений, принятая с самых первых дней эры быстродействующих вычислительных машин,—это представление в *обратном коде*. В этом случае число $-12345\,67890$ записывается в виде

$$87654\,32109. \tag{4}$$

Каждая цифра отрицательного числа $-x$ равна разности между 9 и соответствующей цифрой x. Нетрудно видеть, что для отрицательного числа его обратный код всегда на единицу меньше дополнительного; сложение и вычитание производятся по модулю $10^{10} - 1$, а это означает, что перенос из крайней левой позиции добавляется к крайней правой (см. п. 3.2.1.1). Снова возникают трудности с минус нулем, так как записи $99999\,99999$ и $00000\,00000$ обозначают одно и то же.

Только что изложенные идеи, относящиеся к арифметике по основанию 10, аналогичным образом применимы к арифметике по основанию 2, и мы получаем *двоичные прямой*, *дополнительный* и *обратный коды*. В примерах этой главы машина MIX используется только для работы с представлением в прямом, коде; соответствующие процедуры для дополнительного и обратного кодов обсуждаются, если это оказывается важным, в сопроводительном тексте.

Большинство руководств по вычислительным машинам сообщают, что машинной схемой допускается, чтобы позиционная точка располагалась в фиксированной позиции внутри каждого машинного слова. Это извещение стоит обычно игнорировать; гораздо лучше выучить правила, касающиеся того, где появится позиционная точка в результате выполнения команды, если предположить, что до ее выполнения она расположена на каком-то определенном месте. Например, в случае машины MIX мы могли бы рассматривать наши операнды либо как целые числа с позиционной точкой в крайнем правом положении, либо как правильные дроби с позиционной точкой в крайнем левом положении, либо как некоторые промежуточные между этими двумя крайними вариантами; правила расстановки позиционной точки в каждом результате получаются непосредственно.

Легко видеть, что существует простая связь между записью чисел в системах счисления по основаниям b и b^k:

$$(\ldots a_3a_2a_1a_0.a_{-1}a_{-2}\ldots)_b = (\ldots A_3A_2A_1A_0.A_{-1}A_{-2}\ldots)_{b^k}, \tag{5}$$

где

$$A_j = (a_{kj+k-1} \ldots a_{kj+1} a_{kj})_b;$$

см. упр. 8. Таким образом, мы получаем простой способ переходить "с одного взгляда" от, скажем, двоичной к восьмеричной системе и обратно.

Имеется много интересных вариантов позиционных систем счисления, помимо стандартных b-арных систем, обсуждавшихся до сих пор. Например, мы могли бы рассматривать числа по основанию (-10), так что

$$\begin{aligned}(\ldots a_3 a_2 a_1 a_0 . a_{-1} a_{-2} \ldots)_{-10} = \\ = \ldots + a_3(-10)^3 + a_2(-10)^2 + a_1(-10)^1 + a_0 + \ldots = \\ = \ldots - 1000a_3 + 100a_2 - 10a_1 + a_0 - \frac{1}{10}a_{-1} + \frac{1}{100}a_{-2} - \ldots .\end{aligned}$$

Здесь, как и в обычной десятичной системе, цифры a_k удовлетворяют неравенствам $0 \le a_k \le 9$. Число $12345\,67890$ запишется в такой "нега-десятичной" системе в виде

$$(1\,93755\,73910)_{-10}, \tag{6}$$

так как оно равно как раз $10305070900 - 9070503010$. Интересно отметить, что его знаковое обращение, отрицательное число $-12345\,67890$, записывается в виде

$$(28466\,48290)_{-10}, \tag{7}$$

и в действительности *любое вещественное число, положительное или отрицательное, может быть представлено в системе по основанию* -10 *без знака.*

Системы по отрицательному основанию были упомянуты в литературе впервые, по-видимому, З. Павляком и А. Вакуличем [*Bulletin de l'Academie Polonaise des Sciences,* Classe III, **5** (1957), 233–236; Série des sciences techniques, **7** (1959), 713–721] и Л. Уэйделом [*IRE Transactions,* **EC-6** (1957), 123]. Дальнейшие литературные ссылки можно найти в журналах *IEEE Transactions* [**EC-12** (1963), 274–276] и *Computer Design* [**6** (May, 1967), 52–63]. (Имеются свидетельства того, что идея отрицательного основания возникла независимо сразу у целого ряда лиц по причине растущего интереса к проектированию ЭВМ.)

Дж. Ф. Сонгстер, по предложению Дж. У. Пэттерсона, исследовал системы но основанию -2 в своей магистерской диссертации (Пенсильванский университет, 1956 г.). Системы с отрицательным основанием рассматривал также в 1955 г. Д. Э. Кнут в небольшом машинописном тексте, предназначенном для конкурса "Поиск научных талантов" среди учеников старших классов; там же обсуждалось и дальнейшее обобщение—до комплекснозначных оснований. Тот факт, что все числа могут быть представлены по отрицательному основанию, отмечал в другом контексте несколькими годами ранее Н. Г. де Брёйн [*Publ. Math. Debrecen,* **1** (1950), 232–242, особенно см. стр. 240], однако он не применил эту идею к арифметике.

Выбор основания $2i$ приводит к интересной системе счисления, которую естественно назвать "мнимо-четверичной" (по аналогия с "четверичной"[15], ввиду того что *каждое комплексное число может быть представлено в этой системе при помощи цифр* $0, 1, 2$ *и* 3*, причем тех же цифр, взятых со знаком минус, не требуется).* [См. *CACM,* **3**, (1960), 245–247.] Например,

$$(11210.31)_{2i} = 1 \cdot 16 + 1 \cdot (-8i) + 2 \cdot (-4) + 1 \cdot (2i) + 3 \cdot \left(-\frac{1}{2}i\right) + 1\left(-\frac{1}{4}\right) = 7\frac{3}{4} - 7\frac{1}{2}i.$$

Число $(a_{2n} \ldots a_1 a_0 . a_{-1} \ldots a_{-2k})_{2i}$ равно

$$(a_{2n} \ldots a_2 a_0 . a_{-2} \ldots a_{-2k})_{-4} + 2i(a_{2n-1} \ldots a_3 a_1 . a_{-1} \ldots a_{-2k+1})_{-4},$$

так что перевод числа в мнимо-четверичную форму и обратно сводится к переводу в "нега-четверичную" форму и обратно. Интересное свойство этой системы состоит в том, что она допускает выполнение умножения и деления комплексных чисел целостным образом без раздельного рассмотрения вещественных и мнимых частей. Например, перемножить два числа мы можем в этой системе так же, как и при любом другом основании, используя только несколько иное "правило переноса": в случае если цифра становится больше 4, мы вычитаем 4 и "переносим" -1 на два столбца влево,

[15] В оригинале аналогия "полней": "quaternary"—"quater-imaginary".—*Прим. ред.*

а когда получается отрицательная цифра, мы прибавляем к ней 4 и "переносим" $+1$ на два столбца влево. Разбор нижеследующего примера пояснит, как работает это своеобразное правило переноса:

$$\begin{array}{cccccccccc}
 & & & & 1 & 2 & 2 & 3 & 1 & [9-10i] \\
 & & & & 1 & 2 & 2 & 3 & 1 & [9-10i] \\
\hline
 & & & & 1 & 2 & 2 & 3 & 1 & \\
1 & 0 & 3 & 2 & 0 & 2 & 1 & 3 & & \\
 & & 1 & 3 & 0 & 2 & 2 & & & \\
 & 1 & 3 & 0 & 2 & 2 & & & & \\
1 & 2 & 2 & 3 & 1 & & & & & \\
\hline
0 & 2 & 1 & 3 & 3 & 3 & 1 & 2 & 1 & [-19-180i]
\end{array}$$

Можно построить аналогичную систему по основанию $\sqrt{2}i$, в которой используются лишь цифры 0 и 1, но в этой системе уже для представления самой мнимой единицы i требуется бесконечное непериодическое разложение.

"Бинарную" комплексную систему счисления можно также получить, используя основание $i-1$, предложенное **У. Пенни [*JACM*, 12 (1965), 247–248]:**

$$(\ldots a_4a_3a_2a_1a_0.a_{-1}\ldots)_{i-1} = \ldots - 4a_4 + (2+2i)a_3 - 2ia_2 + (i-1)a_1 + a_0 - \frac{1}{2}(i+1)a_{-1} + \ldots .$$

В этой системе используются только цифры 0 и 1. Один из способов доказать, что каждое комплексное число допускает такое представление, состоит в рассмотрении интересного множества S, изображенного на рис. **1**[16]; это множество состоят по определению из всех точек комплексной плоскости, которые допускают запись в виде $\sum_{k\ge1} a_k(i-1)^{-k}$ для некоторой бесконечной последовательности $a_1, a_2, a_3, \ldots$ нулей и единиц. На рисунке показано, как можно разбить множество S на **256** частей, конгруэнтных $(1/16)S$; заметим, что если множество S повернуть по часовой стрелке на $135°$, то мы увидим, что оно распадается на два примыкающих друг к другу множества, конгруэнтных $(1/\sqrt{2})S$ **(**поскольку $(i-1)S = S \cup (S+1)$**).** По поводу деталей доказательства

Picture: Рис. **1.** Множество S.

того, что S содержит все комплексные числа достаточно малого модуля, см. упр. **18. (**В действительности граница S содержит много мелких "зубцов"; эти зубцы на рис. **1** сглажены.**)**

Быть может, самой изящной из всех является *уравновешенная троичная* система счисления— система по основанию 3, в которой вместо цифр 0, 1, 2 используются "триты"[17] -1, 0, $+1$. Если вместо -1 писать $\overline{1}$, то мы получим следующие примеры чисел, записанных в уравновешенной троичной системе счисления:

Уравновешенная троичная система счисления	Десятичная система счисления
$10\overline{1}$	8
$11\overline{1}0.\overline{11}$	$32\frac{5}{9}$
$\overline{11}10.11$	$-32\frac{5}{9}$
$\overline{11}10.$	-33
$0.11111\ldots$	$\frac{1}{2}$

Уравновешенная троичная система счисления обладает многими приятными свойствами:

a) Переход от числа к противоположному по знаку осуществляется взаимной заменой 1 на $\overline{1}$.
b) Знак числа задается его наиболее значимым ненулевым тритом, и, более общо, мы можем сравнивать любые два числа, используя лексикографический порядок при чтении слева направо, как и в десятичной системе.

[16] С множеством S и тем самым с системой счисления по основанию $i-1$ тесно связаны "кривые дракона", изобретенные и названные так американским физиком Дж. Хейвеем. Замечательные свойства этих кривых исследовал автор настоящей книги Дональд Кнут (см., например, статью Н. Б. Васильева и В. Л. Гутенмахера во 2-м номере журнала "Квант" за 1970 г., 36–46).—*Прим. перев.*

[17] По аналогии с "битами".—*Прим. ред.*

c) Операция округления до ближайшего целого сводится к отбрасыванию дробной части (т. е. всех тритов, стоящих справа от позиционной точки).

Складывать в уравновешенной троичной системе совсем просто, если пользоваться таблицей сложения

$\overline{1}$	$\overline{1}$	$\overline{1}$	$\overline{1}$	$\overline{1}$	$\overline{1}$	$\overline{1}$	$\overline{1}$	$\overline{1}$	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
$\overline{1}$	$\overline{1}$	$\overline{1}$	0	0	0	1	1	1	$\overline{1}$	$\overline{1}$	$\overline{1}$	0	0	0	1	1	1	$\overline{1}$	$\overline{1}$	$\overline{1}$	0	0	0	1	1	1
$\overline{1}$	0	1	$\overline{1}$	0	1	$\overline{1}$	0	1	$\overline{1}$	0	1	$\overline{1}$	0	1	$\overline{1}$	0	1	$\overline{1}$	0	1	$\overline{1}$	0	1	$\overline{1}$	0	1
$\overline{10}$	$\overline{11}$	$\overline{1}$	$\overline{11}$	$\overline{1}$	0	$\overline{1}$	0	1	$\overline{11}$	$\overline{1}$	0	$\overline{1}$	0	1	0	1	$1\overline{1}$	$\overline{1}$	0	1	0	$1\overline{1}$	1	$1\overline{1}$	10	

(Три входных трита—это триты двух наших слагаемых и трю переноса.) Вычитание состоит в переходе к числу, противоположному по знаку, и последующем сложении; умножение также сводится к перемене знака и сложению, как в следующем примере:

			1	$\overline{1}$	0	$\overline{1}$	[17]
			1	$\overline{1}$	0	$\overline{1}$	[17]
			$\overline{1}$	1	0	1	
		$\overline{1}$	1	0	1	0	
	1	$\overline{1}$	0	$\overline{1}$			
0	1	1	$\overline{1}$	$\overline{1}$	0	1	[289]

По поводу деления см. упр. 4.3.1-31.

Один из способов найти представление числа в уравновешенной троичной системе состоит в том, что сначала записывают это число в троичной системе; например,

$$208.3 = (21\,201.022002200220\ldots)_3.$$

(Очень простой способ перехода к троичной системе, пригодный для вычисления вручную, с карандашом и бумагой, описан в упр. 4.4-12.) Далее складываем это число в троичной системе с бесконечным числом $\ldots 11111.11111\ldots$; для нашего примера мы получим

$$(\ldots 11111210012.210121012101\ldots)_3.$$

Наконец, поразрядно вычитаем $\ldots 11111.11111\ldots$, уменьшая на единицу каждую цифру; мы получим

$$208.3 = (10\overline{1}\overline{1}01.10\overline{1}010\overline{1}010\overline{1}0\ldots)_3. \tag{8}$$

Этот процесс, очевидно, можно сделать вполне "законным", если заменить искусственное бесконечное число $\ldots 11111.11111\ldots$ некоторым числом с соответствующим количеством единиц.

Представление чисел в уравновешенной троичной системе неявно присутствует в одной знаменитой математической головоломке, обычно называемой "задачей Башэ о весах", хотя она была сформулирована еще Фибоначчи за четыре столетия до того, как Башэ написал свою книгу. [См. W. Ahrens, Mathematische Unterhaltungen und Spiele, **1**, Leipzig, Teubner, 1910, § 3.4.]

Позиционные системы счисления с отрицательными цифрами были изобретены сэром Джоном Лесли [The philosophy of arithmetic, Edinburgh, 1817; см. стр. 33–34, 54, 64–65, 117, 150] и независимо О. Коши [*Comptes Rendus,* **11** (1840), 789–798], который отмечал, что отрицательные цифры избавляют от необходимости запоминать таблицу умножения дальше 5×5. В "чистом" виде уравновешенная троичная система счисления впервые появилась в статье Леона Лаланна [*Comptes Rendus,* **11** (1840), 903–905], изобретателя механических вычислительных устройств. В течение последующих ста лет после работы Лаланна эта система упоминалась лишь эпизодически, пока в Электротехническом институте Мура в 1945–1946 гг. не стали разрабатывать первые электронные вычислительные устройства; в этот период она серьезно рассматривалась наряду с двоичной системой в качестве возможной замены десятичной системы. Сложность арифметических электронных схем для уравновешенной троичной арифметики не намного выше, чем для двоичной арифметики, а чтобы задать число, в ней требуется лишь $\ln 2/\ln 3 \approx 63\%$ цифровых позиций от того количества, которое нужно в случае двоичной записи. Обсуждение уравновешенной троичной системы см. в журнале *AMM* [**57** (1950), 90–93] и в сборнике "High-speed computing devices" [Engineering Research Associates,

McGraw-Hill, 1950, 287–289]. До сих пор уравновешенная троичная система все еще не нашла серьезного применения, но возможно, что ее симметричность и простая арифметика окажутся в один прекрасный день весьма существенными (когда "флип-флоп" заменится на "флип-флэп-флоп"[18]).

Другое важное обобщение простой позиционной системы—это позиционная система *со смешанными основаниями,* (или *по смешанным основаниям*). Если дана последовательность чисел $\langle b_k \rangle$ (где k могут быть и отрицательными), то мы полагаем по определению

$$\begin{bmatrix} \ldots, & a_3, & a_2, & a_1, & a_0; & a_{-1}, & a_{-2}, & \ldots \\ \ldots, & b_3, & b_2, & b_1, & b_0; & b_{-1}, & b_{-2}, & \ldots \end{bmatrix} = \ldots + a_3 b_2 b_1 b_0 + a_2 b_1 b_0 + a_1 b_0 + a_0 + a_{-1}/b_{-1} + a_{-2}/b_{-1} b_{-2} + \ldots . \tag{9}$$

В простейших системах со смешанными основаниями мы работаем только с целыми числами: мы выбираем в качестве чисел b_0, b_1, b_2, ... целые числа, большие единицы, и рассматриваем только такие числа, которые не содержат позиционной точки, причем число a_k должно принадлежать интервалу $0 \le a_k < b_k$.

Одна из наиболее важных систем со смешанными основаниями—это *факториальная система счисления,* где $b_k = k + 2$. Используя эту систему, мы можем единственным образом представить любое неотрицательное целое число в виде

$$c_n n! + c_{n-1}(n-1)! + \cdots + c_2 2! + c_1, \tag{10}$$

где $0 \le c_k \le k$.

Системы со смешанными основаниями знакомы всем из повседневной жизни; речь идет о единицах мер. Например, величина "три недели, 2 дня, 9 часов, 22 минуты, 57 секунд и 492 миллисекунды" равна

$$\begin{bmatrix} 3, & 2, & 9, & 22, & 57; & 492 \\ & 7, & 24, & 60, & 60; & 1000 \end{bmatrix} \text{ секунд.}$$

В Англии до перехода к десятичной денежной системе величина "10 фунтов, 6 шиллингов, три с половиной пенса" составляла

$$\begin{bmatrix} 10, & 6, & 3; & 1 \\ & 20, & 12; & 2 \end{bmatrix} \text{ пенсов.}$$

Числа по смешанным основаниям можно складывать и вычитать, используя непосредственное обобщение обычных алгоритмов сложения и вычитания, при условии, конечно, что для обоих операндов используется одна и та же система (см. упр. 4.3.1-9). Подобным же образом легко умножать или делить числа по смешанным основаниям на малые целые числа, используя простые обобщения общеизвестных приемов счета при помощи карандаша и бумаги.

В общем виде системы по смешанным основаниям впервые обсуждались Георгом Кантором [*Zeitschrift für Mathematik und Physik,* **14** (1869), 121–128]. Дополнительная информация о таких системах содержится в упр. 26 и 29.

Помимо систем счисления, описанных в этом параграфе, существует несколько других способов представления чисел, которые упоминаются в различных разделах этой серии книг: биномиальная система (упр. 1.2.8-35), система Фибоначчи (упр. 1.2.8-34); фи-система (упр. 1.28-35), модулярное представление (п. 4.3.2), код Грэя (п. 7.2.1) и латинские числа (§ 9.1).

Некоторые вопросы, относящиеся к *иррациональным* основаниям, были исследованы У. Пэрри [*Acta Mathematica,* Acad. Sci. Hung., **11** (1960), 401–416].

Упражнения

1. [15] Выразите числа -10, -9, -8, ..., 8, 9, 10 в системе счисления по основанию -2.

>2. [24] Рассмотрите следующие четыре системы счисления: (a) двоичную (прямой код); (b) негадвоичную (основание -2); (c) уравновешенную троичную; (d) по основанию $b = 1/10$. Используйте каждую из этих систем для представления таких трех чисел: (i) -49, (ii) $-3\frac{1}{7}$ (укажите период); (iii) π (несколько значащих цифр).

3. [20] Выразите $-49 + i$ в мнимо-четверичной системе.

4. [15] Предположим, что в MIX-программе ячейка памяти A содержит число, позиционная точка которого находится между 3-м и 4-м байтами, а ячейка памяти B—число, позиционная точка которого расположена между 2-м и 3-м байтами. (Самый левый байт имеет номер 1.) Где будет располагаться позиционная точка в регистрах A и X после выполнения команд

[18] Flip—щелчок, flap—хлопок, flop—шлепок (*англ.*); flip-flop—принятое в англоязычной литературе название триггера.—*Прим. ред.*

a) 01 LDA A
02 MUL B?

b) 01 LDA A
02 SRAX 5
03 DIV B?

5. **[00]** Объясните, почему представление отрицательного целого числа в обратном десятичном коде всегда на единицу меньше представления в дополнительном коде, если рассматривать эти представления как положительные числа.
6. **[16]** Каковы наибольшие и наименьшие p-разрядные целые числа, которые могут быть представлены в двоичной системе посредством (a) прямого кода, (b) дополнительного кода, (c) обратного кода?
7. **[M20]** В тексте представление в дополнительном десятичном коде определено только для целых чисел, записанных в одном машинном слове. Существует ли способ аналогично определить представление в дополнительном десятичном коде *для всех вещественных чисел,* имеющее "бесконечную точность"? Существует ли подобный способ определить представление в обратном десятичном коде для всех вещественных чисел?
8. **[M10]** Докажите соотношение (5).

>9. **[15]** Переведите следующие восьмеричные числа в шестнадцатеричные (используя шестнадцатеричные цифры 0, 1, ..., F): *12*; *5655*; *2550276*; *76545336*; *3726755*.

10. **[M22]** Обобщите соотношение (5) на случаи систем по смешанным основаниям.
11. **[22]** Используя систему счисления по основанию -2, дайте алгоритм вычисления суммы чисел $(a_n \ldots a_1 a_0)_{-2}$ и $(b_n \ldots b_1 b_0)_{-2}$, получающий ответ $(c_{n+2} \ldots c_1 c_0)_{-2}$.
12. **[23]** Дайте алгоритмы перехода (a) от записи числа в прямом двоичном коде $\pm(a_n \ldots a_0)_2$ к его нега-двоичной записи $(b_{n+1} \ldots b_0)_{-2}$; (b) от нега-двоичной записи $(b_{n+1} \ldots b_0)_{-2}$ к представлению числа в прямом двоичном коде $\pm(a_{n+1} \ldots a_0)_2$.

>13. **[M21]** Существуют числа, которые в десятичной системе счисления имеют два различных бесконечных разложения в десятичную дробь, например $2.3599999\ldots = 2.3600000\ldots$. Единственно ли представление чисел в *нега-десятичной* (по основанию -10) системе счисления или для этого основания также существуют вещественные числа с двумя различными бесконечными разложениями?

14. **[14]** Умножьте $(11321)_{2i}$ на себя в мнимо-четверичной системе, используя описанный в тексте метод.
15. **[M24]** Как выглядят множества S, аналогичные множеству на рис. 1, для нега-десятичной и мнимо-четверичной систем? Другими словами, что представляют собой множества

$$\left\{ \sum_{k\ge 1} a_k(-10)^{-k} \mid 0 \le a_k \le 9, \qquad a_k \text{ целое для всех } k \right\}$$

и

$$\left\{ \sum_{k\ge 1} a_k(2i)^{-k} \mid 0 \le a_k \le 3, \qquad a_k \text{ целое для всех } k \right\}?$$

16. **[M24]** Постройте алгоритм, прибавляющий 1 к $(a_n \ldots a_1 a_0)_{i-1}$ в системе счисления по основанию $i-1$.
17. **[M30]** Может показаться странным, что в качестве основания в системе счисления берется число $i-1$, а не аналогичное, но более простое число $i+1$. Всякое ли комплексное число $a+bi$ с целыми a и b представимо в позиционной системе счисления с цифрами 0 и 1 и основанием $i+1$?
18. **[BM32]** Покажите, что множество S на рис. 1 есть замкнутое множество, содержащее некоторую окрестность начала координат. (Следовательно, любое комплексное число допускает "двоичное" представление по основанию $i-1$.)
19. **[BM42]** Проведите более подробное исследование свойств множества S на рис. 1; например, изучите его границу.
20. **[M22]** Покажите, что любое вещественное число (положительное, отрицательное или нуль) можно представить в десятичной системе счисления при помощи цифр -1, 0, 1, 2, 3, 4, 5, 6, 7, 8 (*без* цифры 9).

>21. **[M22]** (К. Э. Шеннон.) Можно ли произвольное вещественное число (положительное, отрицательное или нуль) представить в "уравновешенной десятичной" системе счисления, т. е. представить в виде $\sum_{k\le n} a_k 10^k$ для некоторого целого n и некоторой последовательности a_n, a_{n-1}, a_{n-2}, ..., где каждое a_k есть одно из десяти чисел $\{ -4\frac{1}{2}, -3\frac{1}{2}, -2\frac{1}{2}, -1\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, 1\frac{1}{2}, 2\frac{1}{2}, 3\frac{1}{2}, 4\frac{1}{2} \}$? (Отметим, что

нуль не входит в число "дозволенных" цифр, однако неявно мы предполагаем, что все цифры a_{n+1}, a_{n+2}, ... суть нули.) Найдите все представления нуля в этой системе и все представления единицы.

22. **[BM25]** Пусть $\alpha = -\sum_{m\geq 1} 10^{-m^2}$. Докажите, что для любого данного $\varepsilon > 0$ и любого вещественного числа x существует такое "десятичное" представление этого числа, что $0 < \left| x - \sum_{0\leq k\leq n} a_k 10^k \right| < \varepsilon$, где каждое из чисел a_k может принимать только три значения: 0, 1 или α. (Отметим, что в этом представлении отрицательные степени 10 не используются!)
23. **[BM30]** Найдите все множества D, состоящие из десяти или меньшего числа неотрицательных вещественных чисел, такие, что (a) $0 \in D$ и (b) все положительные вещественные числа допускают "десятичное" представление $\sum_{k\leq n} a_k 10^k$, где каждое $a_k \in D$.
24. **[BM50]** Найдите все множества D, состоящие из десяти или меньшего числа вещественных чисел, такие, что *любое* неотрицательное вещественное число может быть представлено в виде $\sum_{k\leq n} a_k 10^k$ для некоторого n, где все $a_k \in D$. (Ср. с упр. 20–23.)
25. **[BM25]** (С. А. Кук.) Пусть b, u и v—целые положительные числа, причем $b \geq 2$ и $0 < v < b^m$. Покажите, что представление числа u/v по основанию b не содержит нигде справа от позиционной точки последовательности из m цифр, равных $b-1$. (Согласно общепринятому соглашению, в стандартном представлении по основанию b не допускаются бесконечные последовательности цифр "$b-1$".)

>26. **[BM30]** (Н. С. Мендельсон.) .Пусть $\langle \beta_n \rangle$—последовательность вещественных чисел, определенная для всех целых n, $-\infty < n < +\infty$, такая, что

$$\beta_n < \beta_{n+1}, \quad \lim_{n\to\infty} \beta_n = \infty, \quad \lim_{n\to-\infty} \beta_n = 0,$$

и пусть $\langle c_n \rangle$—произвольная последовательность положительных целых чисел, также определенная для всех целых n, $-\infty < n < +\infty$. Условимся говорить, что число x допускает "обобщенное представление", если существует такое целое n и такая последовательность целых чисел a_n, a_{n-1}, a_{n-2}, ..., что $x = \sum_{k\leq n} a_k \beta_k$, где $a_n \neq 0$, $0 \leq a_k \leq c_k$ и $a_k < c_k$ для бесконечно многих k.

Покажите, что каждое положительное вещественное число x допускает ровно одно обобщенное представление в том и только в том случае, если $\beta_{n+1} = \sum_{k\leq n} c_k \beta_k$ для всех n. (Следовательно, системы по смешанным целочисленным основаниям обладают этим свойством; наиболее общими системами такого типа являются системы по смешанным основаниям, у которых $\beta_1 = (c_0+1)\beta_0$, $\beta_2 = (c_1+1)(c_0+1)\beta_0$, ..., $\beta_{-1} = \beta_0/(c_{-1}+1)$, ...

27. **[M21]** покажите, что любое ненулевое число имеет единственное "знакопеременное двоичное представление" $2^{e_0} - 2^{e_1} + \cdots + (-1)^k 2^{e_k}$, где $e_0 < e_1 < \ldots < e_k$.

>28. **[M24]** Покажите, что каждое ненулевое комплексное число вида $a + bi$, где числа a и b целые, обладает единственным "вращательным двоичным представлением"

$$(1+i)^{e_0} + i(1+i)^{e_1} - (1+i)^{e_2} - i(1+i)^{e_3} + \cdots + i^k(1+i)^{e_k},$$

где $e_0 < e_1 < \ldots < e_k$. (Ср. с упр. 27.)

29. **[M35]** (Н. де Брёйн.) Пусть S_0, S_1, S_2, ... —множества неотрицательных целых чисел; мы скажем, что совокупность S_0, S_1, S_2, ... обладает свойством B, если любое неотрицательное целое число n может быть представлено в виде

$$n = s_0 + s_1 + s_2 + \ldots, \qquad s_i \in S_j$$

ровно одним способом. (Свойство B означает, в частности, что $0 \in S_j$ для всех f, ибо $n = 0$ может быть представлено только как $0 + 0 + 0 + \ldots$.) Любая система счисления со смешанными основаниями b_0, b_1, b_2, ... доставляет пример совокупности множеств, удовлетворяющих свойству B, если положить $S_j = \{0, q_j, \ldots, (b_j - 1)q_j\}$, где $q_j = b_0 b_1 \ldots b_{j-1}$; в этом случае представление $n = s_0 + s_1 + s_2 + \ldots$ очевидным образом соответствует представлению (9) этого числа по смешанным основаниям. Далее, если совокупность S_0, S_1, S_2, ... обладает свойством B, то, каково бы ни было разбиение A_0, A_1, A_2, ... множества неотрицательных целых чисел (т. е. $A_0 \cup A_1 \cup A_2 \cup \ldots = \{0, 1, 2, \ldots\}$ и $A_i \cap A_j = \emptyset$ при $i \neq j$; некоторые из множеств A_j могут быть пустыми), этим свойством обладает и полученная из нее "стягиванием"[19]совокупность T_0, T_1, T_2, ..., где множество T_j состоит из всех сумм вида $\sum_{i\in A_j} S_i$, взятых по всевозможным выборкам элементов $s_i \in S_i$.

[19] В оригинале "collapsing".—*Прим. ред.*

Докажите, что *любая* совокупность T_0, T_1, T_2, ..., удовлетворяющая свойству B, может быть получена стягиванием некоторой совокупности S_0, S_1, S_2, ..., соответствующей системе счисления по смешанным основаниям.

30. [M39] (Н. де Брёйн.) Пример системы счисления по основанию -2 показывает, что любое целое число (положительное, отрицательное или нуль) имеет единственное представление в виде

$$(-2)^{e_1} + (-2)^{e_2} + \cdots + (-2)^{e_t}, \qquad e_1 > e_2 > \ldots > e_t \geq 0, \quad t \geq 0.$$

Цель данного упражнения состоит в исследовании некоторых обобщений этого феномена.

a) Пусть последовательность целых чисел b_0, b_1, b_2, ... обладает тем свойством, что любое целое число n допускает единственное представление в виде

$$n = b_{e_1} + b_{e_2} + \cdots + b_{e_t}, \qquad e_1 > e_2 > \ldots > e_t \geq 0, \quad t \geq 0$$

(такая последовательность $\langle b_n \rangle$ называется "бинарным базисом"). Покажите, что найдется такое значение индекса j, что b_j нечетно, а для всех $k \neq j$ числа b_k четны.

b) Докажите, что бинарный базис $\langle b_n \rangle$ может быть всегда переупорядочен в последовательность вида d_0, $2d_1$, $4d_2$, $\ldots = \langle 2^n d_n \rangle$, где каждое из чисел d_k нечетно.

c) Докажите, что если каждое из чисел d_0, d_1, d_2, ... из пункта b) равно ± 1, то последовательность $\langle b_n \rangle$ образует бинарный базис тогда и только тогда, когда существует бесконечно много d_j, равных $+1$, и бесконечно много d_j, равных -1.

d) Докажите, что последовательность 7, $-13 \cdot 2$, $7 \cdot 2^2$, $-13 \cdot 2^3$, ..., $7 \cdot 2^{2k}$, $-13 \cdot 2^{2k+1}$, ... является бинарным базисом, и найдите представление числа $n = 1$.

>31. [M35] Одно обобщение представления чисел в дополнительном двоичном коде, известное под названием "2-адических чисел", было изобретено К. Гензелем около 1900 г. (В действительности Гензель изобрел *p-адические числа* для любого простого числа p.) А именно 2-адическое число можно рассматривать как двоичное число

$$u = (\ldots u_3 u_2 u_1 u_0 . u_{-1} \ldots u_{-n})_2,$$

представление которого продолжается бесконечно далеко влево и лишь на конечное число знаков вправо от разделительной точки. Сложение, вычитание и умножение 2-адических чисел выполняются в соответствии с обычными арифметическими процедурами, которые в принципе допускают возможность неограниченного продолжения влево. Например:

$$\begin{aligned} 7 &= (\ldots 000000000000111)_2, & 1/7 &= (\ldots 110110110110111)_2, \\ -7 &= (\ldots 111111111111001)_2, & -1/7 &= (\ldots 001001001001001)_2, \\ 7/4 &= (\ldots 000000000000001.11)_2, & 1/10 &= (\ldots 110011001100110.1)_2, \\ \sqrt{-7} &= (\ldots 100000010110101)_2 \text{ или } (\ldots 011111101001011)_2. \end{aligned}$$

Здесь 7—обычное целое число семь в двоичном представлении, а -7 есть его дополнительный код (неограниченно продолженный влево); легко проверить, что обычная процедура сложения двоичных чисел даст нам $-7 + 7 = (\ldots 00000)_2 = 0$, если выполнение этой процедуры продолжать неограниченно долго. Значения $1/7$ и $-1/7$ представляют собой единственные 2-адические числа, которые после формального умножения на 7 дают соответственно $+1$ и -1. Значения $7/4$ и $1/10$ служат примерами 2-адических чисел, не являющихся 2-адическими "целыми", так как они имеют ненулевые биты справа от разделительной точки. Приведенные два значения $\sqrt{-7}$, получающихся одно из другого переменой знака, суть 2-адические числа, которые после формального возведения в квадрат дают $(\ldots 111111111111001)_2$.

a) Докажите, что любое 2-адическое число u можно разделить на произвольное ненулевое 2-адическое число v в том смысле, что существует единственное 2-адическое число w, удовлетворяющее равенству $u = vw$. (Следовательно, множество 2-адических чисел образует поле; см. п. 4.6.1.)

b) Докажите, что 2-адическое представление рационального числа $1/(2n+1)$, где n—целое положительное число, можно получить следующим образом. Сначала находим обычное двоичное разложение числа $1/(2n+1)$, которое имеет вид "периодической дроби" $(0.\alpha\alpha\alpha\ldots)_2$, где α—некоторая строка из нулей и единиц. Тогда 2-адическим представлением числа $-1/(2n+1)$ будет $(\ldots\alpha\alpha\alpha)_2$.

c) Докажите, что 2-адическое представление числа u периодично (т. е. $u_{N+\lambda} = u_N$ для всех больших N при некотором $\lambda \geq 1$) тогда и только тогда, когда u рационально (т.е. $u = m/n$ для некоторых целых чисел m и n).

d) Докажите, что если n—целое число, то $\sqrt{n}$ является 2-адическим числом в том и только том случае, когда $n \bmod 2^{2k+3} = 2^{2k}$ для некоторого неотрицательного целого k. (Таким образом, либо $n \bmod 8 = 1$, либо $n \bmod 32 = 4$ и т. д.)

4.2. АРИФМЕТИКА ЧИСЕЛ С ПЛАВАЮЩЕЙ ТОЧКОЙ

4.2.1. Вычисления с однократной точностью

В этом параграфе мы рассмотрим основные принципы выполнения арифметических действий над числами с "плавающей точкой" и проанализируем внутренний механизм таких вычислений. Вероятно, у многих читателей этот предмет не вызовет слишком большого интереса либо по той причине, что вычислительные машины, на которых они работают, имеют встроенные команды операций над числами с плавающей точкой, либо потому, что производитель снабдил их ЭВМ нужными подпрограммами. Но в действительности материал этого параграфа не следует считать относящимся исключительно к компетенции инженеров-конструкторов ЭВМ или узкого круга лиц, которые пишут библиотечные подпрограммы для новых машин; *каждый* хороший программист должен иметь представление о том, что происходит при выполнении элементарных шагов арифметических операций над числами с плавающей точкой. Предмет этот совсем не так тривиален, как принято считать; в нем удивительно много интересного.

А. Обозначение чисел с плавающей точкой. В § 4.1 мы обсудили обозначения для чисел с "фиксированной точкой"; в этом случае программист знает, где положено находиться разделительной точке в тех числах, с которыми он работает. Для многих целей при выполнении программы значительно более удобно сделать положение разделительной точки динамической переменной—иными словами, сделать ее "плавающей"—и связать с каждым числом указание о положении соответствующей разделительной точки. Эта идея уже давно использовалась в научных вычислениях, в особенности для представления очень больших чисел типа числа Авогадро $N = 6.02250 \times 10^{23}$ или очень малых чисел типа постоянной Планка $\hbar = 1.0545 \times 10{-}27$ эрг $\cdot$ с.

В этом пункте мы будем иметь дело с *p-разрядными числами с плавающей точкой по основанию b с избытком q*. Такое число представляется как пара величин (e, f), которой отвечает значение

$$(e, f) = f \times b^{e-q}. \tag{1}$$

Здесь e—целое число, изменяющееся в соответствующем интервале значений, и f—дробное число со знаком. Условимся, что

$$|f| < 1,$$

иными словами, разделительная точка в позиционном представлении f находится в крайнем левом положении. Более точно, соглашение о том, что мы имеем дело с p-разрядными числами, означает, что $b^p f$—целое число и

$$-b^p < b^p f < b^p. \tag{2}$$

Термин "двоичный" будет означать, как всегда, что $b = 2$, "десятичный"—что $b = 10$ и т. д. Используя 8-разрядные десятичные числа с плавающей точкой с избытком 50, мы можем, например, написать:

$$\begin{aligned} &\text{число Авогадро } N = (74, +.60225000); \\ &\text{постоянная Планка } \hbar = (24, +.10545000)^{20}. \end{aligned} \tag{3}$$

Две компоненты e и f числа с плавающей точкой называются его *показателем* и *дробной частью* соответственно. (Иногда для этой цели используются и другие названия, в особенности "характеристика" и "мантисса"; однако использование слова "мантисса" для обозначения дробной части приводит к путанице в терминологии, так как этот термин употребляется совсем в другом смысле в теории логарифмов, а кроме того, английское слово mantissa[21] означает "мало дающее добавление".)

В этом пункте мы почти целиком сосредоточим свое внимание на представлении дробной части f в прямом коде, так как представление чисел с плавающей точкой в дополнительном коде не обладает многими желательными свойствами (см. п. 4.2.2).

---

[20] У постоянной Планка шестая значащая цифра неизвестна, поэтому. называть это число постоянной Планка несколько рискованно.—*Прим. ред.*

[21] В вышедшем из употребления значении.—*Прим. ред.*

Число (e, f) с плавающей точкой называется *нормализованным*, если наиболее значимая цифра в представлении f отлична от нуля, так что

$$1/b \le |f| < 1, \tag{4}$$

либо если $f = 0$, а e принимает наименьшее возможное значение. Чтобы установить, какое из двух нормализованных чисел с плавающей точкой имеет большую величину, достаточно сравнить их показатели, и только если эти показатели равны, нужно привлечь к рассмотрению и дробные части.

В нашей машине MIX числа с плавающей точкой имеют вид

$\pm$	e	f	f	f	f

Это—представление с плавающей точкой по основанию b с избытком q, с четырьмя значащими "цифрами", где b есть размер байта (например, $b = 64$ или $b = 100$) и q равняется $\lceil \frac{1}{2} b \rceil$. Дробная часть равна $\pm ffff$, а показатель e заключен в интервале $0 \le e < b$. Это внутреннее представление—типичный образец тех соглашений, которые приняты в большинстве существующих ЭВМ, хотя основание b здесь гораздо больше, чем обычно используемые.

B. Нормализованные вычисления. Большинство ныне применяемых стандартных программ работают почти исключительно с нормализованными числами: входные значения для подпрограмм предполагаются нормализованными и значения на выходе всегда нормализуются.

Рассмотрим теперь арифметические операция над числами с плавающей точкой подробнее. Одновременно мы сможем изучать структуру подпрограмм, реализующих эти операции (в предположении, что в нашем распоряжении имеется ЭВМ без схемной реализации действий над числами с плавающей точкой).

Стандартные подпрограммы для арифметических действий над числами с плавающей точкой, когда их пишут на машинном языке, в очень большой степени зависят от конкретной машины и используют многие крайне специфические особенности этой машины. Именно поэтому так мало сходства между двумя подпрограммами, скажем, сложения чисел с плавающей точкой, написанными для разных машин. Все же внимательное рассмотрение большого числа подпрограмм как для двоичных, там и для десятичных машин показывает, что в действительности эти программы имеют много общего, и вполне возможно обсуждение этой темы на машинно-независимом уровне.

Первый (и наиболее трудный!) из алгоритмов, обсуждаемых в этом пункте,—это процедура сложения чисел с плавающей точкой:

$$(e_u, f_u) \oplus (e_v, f_v) = (e_w, f_w). \tag{6}$$

Замечание. Ввиду того что арифметические действия над числами с плавающей точкой являются по самому существу дела приближенными, а не точными, для обозначения операций сложения, вычитания, умножения и деления с плавающей точкой мы используем символы

$$\oplus, \quad \ominus, \quad \otimes, \quad \oslash,$$

с тем чтобы отличать приближенные операции от точных. Идея, лежащая в основе сложения с плавающей точкой, довольна проста: в предположении, что $e_u \ge e_v$, мы берем $e_w = e_u$, $f_w = f_u + f_v/b^{e_u - e_v}$ (таким образом, мы выравниваем положение позиционных точек, чтобы сложение имело смысл), а затем нормализуем результат. Может возникнуть несколько ситуаций, которые делают выполнение этого процесса нетривиальным; более точное описание метода дается следующим алгоритмом.

Алгоритм A. *(Сложение чисел с плавающей точкой)* . Для заданных p-разрядных нормализованных чисел с плавающей точкой $u = (e_u, f_u)$ и $v = (e_v, f_v)$ по основанию b с избытком q строится сумма $w = u \oplus v$. Этот же самый алгоритм можно использовать для вычитания чисел с плавающей точкой, если v заменить на $-v$. Основание b предполагается четным.

A1 [Распаковать.] Выделить показатель и дробную часть в представлениях для u и v.

A2 [Обеспечить справедливость допущения $e_u \ge e_v$.] Если $e_u < e_v$, поменять местами u и v. (Во многих случаях удобнее совместить шаг **A2** с шагом **A1** или с каким-нибудь из последующих шагов.)

A3 [Присвоить значение e_w.] Установить $e_w \leftarrow e_u$.

A4 [Проверить $e_u - e_v$.] Если $e_u - e_v \ge p + 2$ (большая разница в показателях), установить $f_w \leftarrow f_u$ и перейти в шаг **A7**. (Так как мы предполагаем, что u нормализовано, то можно было бы здесь

и закончить выполнение алгоритма, но часто оказывается полезным иметь в распоряжении возможность нормализовать какое-либо, возможно ненормализованное, число, сложив его с нулем.)

A5 [Сдвинуть шкалу вправо.] Сдвинуть f_v вправо на $e_u - e_v$ позиций, т. е. разделить f_v на $b^{e_u-e_v}$. *Замечание.* Величина сдвига может доходить до $p+1$ разрядов, так что следующий шаг (сложение дробных частей f_u и f_v) потребует тем самым наличия аккумулятора, способного хранить $2p+1$ цифр по основанию b справа от позиционной точки. Если такого

Picture: Рис. 2. Сложение чисел с плавающей точкой.

вместительного аккумулятора нет, можно потребовать сдвига до $p+2$ разрядов, но с соответствующими предосторожностями; подробности обсуждаются в упр. 5.

A6 [Сложить.] Установить $f_w \leftarrow f_u + f_v$.

A7 [Нормализовать.] (В этот момент (e_w, f_w) представляет сумму u и v, но f_w может содержать более чем p цифр и может быть больше единицы или меньше $1/b$). Выполнить описываемый ниже алгоритм N, который нормализует и округлит (e_w, f_w) до окончательного ответа. ∎

Алгоритм N. *(Нормализация)* "Сырой" показатель e и "сырая" дробная часть f приводятся к нормализованному виду, с округлением, если нужно, до p разрядов. В этом алгоритме предполагается, что $|f| < b$ и что число b четно.

N1 [Проверить f.] Если $|f| \geq 1$ ("переполнение дробной части"), перейти к выполнению шага **N4**. Если $f = 0$, установить e_w равным его наименьшему значению и перейти в шаг **N7**.

N2 [f нормализовано?] Если $|f| \geq 1/b$, перейти к выполнению шага **N5**.

N3 [Сдвинуть шкалу влево.] Сдвинуть f на один разряд влево (т. е. умножить на b) и уменьшить e на 1. Возвратиться в шаг **N2**.

N4 [Сдвинуть шкалу вправо.] Сдвинуть f вправо на один разряд (т. е. разделить на b) и увеличить e на 1.

N5 [Округлить.] Округлить f до p разрядов. (Это следует понимать в том смысле, что $f \leftarrow b^{-p}\lfloor b^p f + 1/2 \rfloor$, если $f > 0$0, и $f \leftarrow b^{-p}\lceil b^p f - 1/2 \rceil$, если $f < 0$; можно использовать и другие правила округления, но это общее определение, по-видимому, более удачно вписывается в развиваемую дальше в этой главе теорию.) Важно заметить, что эта операция округления может привести к равенству $|f| = 1$ ("переполнение при округлении"); в таком случае следует вернуться в шаг **N4**.

N6 [Проверить e.] Если показатель e слишком велик, т. е. больше допустимой границы, то это воспринимается как сигнал о *переполнении показателя.* Если e слишком мал, то это воспринимается как сигнал об *исчезновении показателя.* (См. дальнейшее обсуждение вопроса ниже; эти ситуации интерпретируются обычно как сигнал об ошибке в том смысле, что результат не может быть представлен в виде нормализованного числа с плавающей точкой из требуемого интервала значений.)

N7 [Упаковать.] Объединить показатель e и дробную часть f для выдачи искомого представления. ∎

Несколько простых примеров сложения чисел с плавающей точкой дано в упр. 4.

Приводимые ниже MIX-подпрограммы для сложения и вычитания чисел, имеющих форму (5), служат примером того, как алгоритмы A и N могут быть реализованы в виде программ для ЭВМ. Эти подпрограммы извлекают одно входное значение u по символическому адресу ACC, а другое входное значение v

Picture: Рис. 3. Нормализация числа (e, f).

извлекается из регистра A при входе в подпрограмму. Результат w появляется одновременно в регистре A и поле ACC. Таким образом, последовательность команд

```
LDA A; ADD B; SUB C; STA D,
```
(7)

работающих с числами с фиксированной точкой, соответствовала бы такой последовательности команд, работающих с числами с плавающей точкой:

```
LDA A, STA ACC; LDA B, JMP FADD; LDA C, JMP FSUB; STA D.
```
(8)

Программа A. *(Сложение, вычитание и нормализация)* . Следующая программа представляет собой подпрограмму, реализующую алгоритм A, причем она построена таким образом, чтобы нормализующий фрагмент мог быть использован другими подпрограммами, которые появятся в этом пункте

в дальнейшем. Как в этой программе, так и во многих других программах этой главы идентификатор OFLO именует подпрограмму, которая печатает сообщение о том, что индикатор переполнения машины MIX внезапно пришел в состояние "включено".

```
EXP  EQU   1:1     Определение поля показателя.
FSUB STA   TEMP    Подпрограмма вычитания величин с плавающей точкой.
     LDAN  TEMP    Изменить знак операнда.
FADD STJ   EXITF   Подпрограмма сложения величин с плавающей точкой:
     JOV   OFLO    Убедиться в том, что переполнение отсутствует.
```

***** Разрыв текста (пропущены стр. 231-240) *****

Большинство публикаций о деталях программ для работы в системе с плавающей точкой рассеяны по "техническим памятным запискам", распространяемым различными производителями ЭВМ, но случались и публикации этих программ в открытой литературе. Помимо приведенных выше работ см. R. H. Stark, D. B. McMillan, *Math. Comp.,* **5** (1951), 86–92, где описана программа для реализации на панельно-штекерном устройстве; D. McCracken, Digital Computer Programming, New York, Wiley, 1957, 121–131; J. W. Carr III, *CACM,* **2** (May, 1959), 10–15; W. G. Wadey, *JACM,* **7** (1960), 129–139; D. E. Knuth, *JACM,* **8** (1961), 119–128; O. Kesner, *CACM,* **5** (1962), 269–271; F. P. Brooks, K. E. Iverson, Automatic data processing, New York, Wiley, 1963, 184–199. Обсуждение арифметических операций в системе с плавающей точкой с точки зрения инженера-электронщика можно найти в статье С. Г. Кэмпбелла "Floating-point operation" в сборнике "Planning a Computer System", [ed. by W. Buchholz, New York, McGrow-Hill, 1962,92–121]. Дополнительные ссылки на литературу в п. 4.2.2.

Упражнения

1. [10] Как будут выглядеть число Авогадро и постоянная Планка, если их представить в виде четырехразрядных чисел с плавающей точкой по основанию 100 с избытком 50? (Именно таково было бы представление в машине MIX (как в (5)), если бы размер байта равнялся 100.)
2. [12] Предположим, что показатель e лежит в интервале $0 \le e \le E$; каковы наибольшее и наименьшее положительные значения, которые могут быть записаны как p-разрядные числа с плавающей точкой по основанию b с избытком q? Каковы наибольшее и наименьшее положительные значения, которые могут быть представлены в виде *нормализованных* таких чисел?
3. [20] Покажите, что если мы работаем с нормализованными двоичными числами с плавающей точкой, то существует способ немного увеличить точность без увеличения объема используемой памяти: p-разрядную дробную часть можно представлять при помощи всего лишь $p-1$ разрядов машинного слова, если чуть-чуть уменьшить интервал значений показателя.

>4. [15] Пусть $b = 10$, $p = 8$. Какой результат даст алгоритм A для операции $(50, +.98765432) \oplus (49, +.33333333)$? Для операции $(53, -.99987654) \oplus (54, +.10000000)$? Для операции $(45, -.50000001) \oplus (54, +.10000000)$?

5. [M23] Докажите, что между шагами A5 и A6 алгоритма A можно, не изменяя результата, поместить следующую операцию: если f_u и f_v имеют одинаковый знак, то заменить f_v на $\operatorname{sign}(f_v) b^{-p-2} \lfloor b^{p+2} |f_v| \rfloor$; если знаки f_u и f_v противоположны, то заменить f_v на $\operatorname{sign}(f_v) b^{-p-2} \lceil b^{p+2} |f_v| \rceil$. (Эффект этой операции состоит в "урезании" f_v до $p+2$ разрядов, чтобы минимизировать длину регистра, необходимую для выполнения сложения в шаге A6.)
6. [22] Удачна ли идея заменить строки 38–40 программы A одной-единственной командой "SLC 5"?
7. [M21] Какие изменения необходимо произвести в алгоритме A, чтобы он обеспечивал выдачу правильно округленного нормализованного результата, даже если входные данные не нормализованы? (Слова "правильно округленный" означают, что результат имеет наибольшую возможную точность в p разрядах в предположении, что входные данные u и v в точности равны $f_u \times b^{e_u - q}$ и $f_v \times b^{e_v - q}$ соответственно, хотя, быть может, и не нормализованы. В частности, u или v могут иметь нулевую дробную часть с очень большим показателем, и такой операнд в контексте данного упражнения воспринимался бы как равный нулю.)
8. [25] Приведите примеры входных значений, для которых подпрограмма FADD в программе A не обеспечивает получения "правильно округленного нормализованного ответа" в смысле упр. 7.
9. [M24] (У. Кахан.) Предположим, что исчезновение показателя приводит к присвоению результату значения нуль без какого-либо указания об ошибке. Используя восьмиразрядные десятичные числа с плавающей точкой с избытком нуль и показателем e в интервале $-50 \le e < 50$, найдите такие положительные значения a, b, c, d и y, для которых выполняются соотношения (11).
10. [M15] Приведите пример нормализованных восьмиразрядных десятичных чисел с плавающей точкой u и v, в процессе сложения которых происходит переполнение при округлении.

>11. [M20] Дайте пример нормализованных восьмиразрядных десятичных чисел с плавающей точкой u и v, в процессе умножения которых происходит переполнение при округлении.

12. [M25] Докажите, что переполнение при округлении не может происходить в ходе выполнения фазы нормализации при делении чисел с плавающей точкой.

>13. [M23] Мистер Смышлёный следующим образом модифицировал прием, приводящий к (13) с тем, чтобы производить *урезание* положительных чисел u с плавающей точкой до целых чисел $\lfloor u \rfloor$ с фиксированной точкой в предположении, что $0 < u < b^3$:

```
FSUB ONEHALF;  FADD FUDGE;  SUB FUDGE.
```

(Здесь ONEHALF—представление $1/2$ в виде числа с плавающей точкой.) Его идея основывалась на том наблюдении, что $u-1/2$ округляется до $\lfloor u \rfloor$. Однако позже он обнаружил, что эта идея работает не всегда. В чем была ошибка? Нет ли другого метода, которым бы он мог воспользоваться?

14. [23] Напишите MIX-подпрограмму, которая, будучи применена к произвольному числу с плавающей точкой, расположенному в регистре A и не обязательно нормализованному, перерабатывает его в ближайшее целое число, представленное в форме с фиксированной точкой (или устанавливает, что это число слишком велико по абсолютной величине, чтобы такое преобразование было возможно).

>15. [28] Напишите MIX-подпрограмму, которую можно было бы использовать в сочетании с другими подпрограммами этого параграфа для вычисления $u \pmod 1$, т. е. для определения $u - \lfloor u \rfloor$, в случае когда число u задано в форме с плавающей точкой. Нужно, чтобы получился правильно округленный результат (в смысле упр. 7); следовательно, если u—очень малое отрицательное число, то $u \pmod 1$ будет округлено таким образом, что результат окажется равным 1 (хотя по определению $u \pmod 1$ есть вещественное число, всегда *меньшее* единицы).

16. [BM21] (Роберт Л. Смит.) Разработайте алгоритм вычисления вещественной и мнимой частей комплексного числа $(a + bi)/(c + di)$ по заданным в форме с плавающей точкой вещественным значениям a, b, c и d. Обойдитесь без вычисления $c^2 + d^2$, так как при этом может возникнуть переполнение, если $|c|$ или $|d|$ примерно равно квадратному корню из максимального допустимого в форме с плавающей точкой значения.

17. [40] (Джон Коки.) Рассмотрите следующую идею расширения интервала изменения чисел с плавающей точкой: используется однословное представление, при котором точность дробной части уменьшается, когда величина показателя возрастает.

4.2.2. Точность выполнения арифметических действий в системе с плавающей точкой

Вычисления над числами в форме с плавающей точкой неточны по самой своей природе, и нетрудно столь неудачно проводить их, чтобы вычисленные ответы почти целиком состояли из "шума". Одна из главных проблем численного анализа состоит в определении степени точности результатов тех или иных численных методов; сюда включается и проблема "степени доверия": мы не знаем, до какой степени можно верить результатам вычислений на ЭВМ. Пользователи-новички решают эту проблему, доверяя компьютеру как непогрешимому авторитету; они склонны считать, что все цифры напечатанного ответа являются значащими. У пользователей ЭВМ, лишенных иллюзий, подход прямо противоположный: они неизменно опасаются, что их ответы почти бессмысленны. Многие из серьезных математиков пытались строго проанализировать последовательность операций с плавающей точкой, но, обнаружив, что задача слишком трудна, кончали тем, что удовлетворялись правдоподобными рассуждениями.

Полное исследование методов анализа ошибок выходит, разумеется, за рамки настоящей книги, однако некоторые из характеристик ошибок, возникающих при вычислениях в системе с плавающей точкой, мы в этом пункте рассмотрим. Наша цель—выяснить, как выполнять операции с плавающей точкой таким образом, чтобы облегчить, насколько это возможно, разумный анализ величины ошибки.

Простой (но достаточно полезный) способ охарактеризовать поведение операций плавающей арифметики основан на понятии "значащих цифр" или *относительной ошибки*. Если мы представляем точное вещественное число x в ЭВМ посредством, приближения $\widehat{x} = x(1 + \varepsilon)$, то величина $\varepsilon = (\widehat{x} - x)/x$ называется относительной ошибкой приближения. Грубо говоря, в системе с плавающей точкой операции умножения и деления не слишком увеличивают относительную ошибку, но вычитание почти равных величин (и сложение $u \oplus v$, где u почти равно $-v$) может увеличить ее значительно. Итак, у нас есть общее "эмпирическое" правило: существенной потери точности можно ожидать от сложения и вычитания указанного вида, но не от умножения и деления.

Одним из следствий о возможной ненадежности сложения в системе с плавающей точкой является нарушение закона ассоциативности:

$$(u \oplus v) \oplus w \neq u \oplus (v \oplus w) \qquad \text{для некоторых } u, v, w. \tag{1}$$

Например:

$$\begin{aligned}(11111113. \oplus -11111111.) \oplus 7.5111111 &= 2.0000000 \oplus 7.5111111\\ &= 9.5111111;\\ 11111113. \oplus (-11111111. \oplus 7.5111111) &= 11111113. \oplus -11111103.\\ &= 10.000000.\end{aligned}$$

(Все примеры этого пункта приводятся в восьмиразрядной десятичной плавающей системе с представлением показателей посредством прямого указания места расположения десятичной точки. Напомним, что, как и в п. 4.2.1, символы $\oplus$, $\ominus$, $\otimes$, $\oslash$ используются для обозначения операций в системе с плавающей точкой, соответствующих точным операциям $+$, $-$, $\times$, $/$.)

В свете возможной утраты закона ассоциативности приведенное в начале этой главы замечание госпожи Ла Туш [взятое из *Math. Gazette,* **12** (1924), 95], если его отнести к арифметике в системе с плавающей точкой, несет в себе большую долю здравого смысла. Математические обозначения типа "$a_1 + a_2 + a_3$" или "$\sum_{1\le k\le n} a_k$" по самому своему существу основаны на предположении об ассоциативности, так что программист должен быть особенно бдителен на тот счет, не предполагает ли он неявно справедливость закона ассоциативности.

А. Аксиоматический подход. Хотя закон ассоциативности и не верен, закон коммутативности

$$u \oplus v = v \oplus u \tag{2}$$

должен выполняться, и этот закон может служить серьезным концептуальным подспорьем при программировании и при анализе программ. Это соображение подсказывает нам, что следует искать наиболее существенные законы, которым удовлетворяют операции $\oplus$, $\ominus$, $\otimes$ и $\oslash$; далее, не будет безрассудным такое высказывание: *программы арифметических операций в системе с плавающей точкой следует составлять таким образом, чтобы сохранялось максимально возможное число стандартных математических законов.*

Рассмотрим теперь некоторые из других основных законов, которые сохраняются для нормализованных операций с плавающей точкой, описанных в предыдущем пункте. Прежде всего мы имеем

$$u \ominus v = u \oplus -v; \tag{3}$$

$$-(u \oplus v) = -u \oplus -v; \tag{4}$$

$$u \oplus v = 0 \text{ тогда и только тогда, когда } v = -u; \tag{5}$$

$$u \oplus 0 = u. \tag{6}$$

Из этих законов можно получать и дальнейшие тождества; например,

$$u \ominus v = -(v \ominus u) \tag{7}$$

(см. упр. 1). Эти законы *не* выполнялись бы, если бы в системе с плавающей точкой для дробных частей операндов вместо прямого кода использовался дополнительный код; см. упр. 11. С этой точки зрения прямой код для представления чисел с плавающей точкой имеет некоторое теоретическое преимущество.

Тождества (2)–(6) следуют непосредственно из алгоритмов п. 4.2.1. Следующее правило чуть менее очевидно:

$$\text{если } u \le v, \text{ то } u \oplus w \le v \oplus w. \tag{8}$$

Чтобы доказать его, положим по определению $\operatorname{round}(x,p) =$ "x, округленное до p разрядов" $=$

$$\begin{cases} b^{e-p}\lfloor b^{p-e}x + \frac{1}{2}\rfloor, & \text{если } b^{e-1} \le x < b^e, \\ 0, & \text{если } x = 0, \\ b^{e-p}\lceil b^{p-e}x - \frac{1}{2}\rceil, & \text{если } b^{e-1} \le -x < b^e, \end{cases} \tag{9}$$

где x—вещественное число, а p—целое положительное. (Ср. с алгоритмом 4.2.1N, шаг N5.) Заметим, что справедливы равенства

$$\operatorname{round}(-x,p) = -\operatorname{round}(x,p); \qquad \operatorname{round}(bx,p) = b\operatorname{round}(x,p); \tag{10}$$

и, далее, в силу определений п. 4.2.1, имеют место важные соотношения

$$u \oplus v = \operatorname{round}(u+v,p), \tag{11}$$

$$u \ominus v = \operatorname{round}(u-v,p), \tag{12}$$

$$u \otimes v = \operatorname{round}(u\times v,p), \tag{13}$$

$$u \oslash v = \operatorname{round}(u/v,p) \tag{14}$$

при условии, что не происходит переполнения или исчезновения показателей, т. е. при условии, что числа $u+v$, $u-v$, $u\times v$ и u/v лежат в допустимых интервалах. Пользуясь этим наблюдением, можно дать простое доказательство всех перечисленных выше тождеств, включая и тождество (8), являющееся следствием того факта, что функция $\mathrm{round}(x,p)$ не убывает при возрастании x.

Мы можем теперь выписать еще несколько тождеств, вытекающих из указанных выше соотношений:

$$u\otimes v=v\otimes u,(-u)\otimes v=-(u\otimes v),1\otimes v=v;$$
$$u\otimes v=0 \text{ тогда и только тогда, когда } u=0 \text{ или } v=0;$$
$$(-u)\oslash v=u\oslash(-v)=-(u\oslash v),0\oslash v=0,$$
$$u\oslash 1=u,u\oslash u=1;$$
$$\text{если } u\le v,\ w>0, \text{ то } u\otimes w\le v\otimes w,\ u\oslash w\le v\oslash w,\ w\oslash u\ge w\oslash v.$$

Итак, если операции в системе с плавающей точкой установлены в соответствии с определенными условиями, то, несмотря на неточность самих операций, сохраняется немалая регулярность.

В приведенной выше коллекции тождеств, разумеется, все же бросается в глаза отсутствие нескольких известных законов алгебры; закон ассоциативности для умножения в системе с плавающей точкой выполняется не вполне точно, как это будет видно из упр. 3, что же касается закона дистрибутивности, связывающего операции $\otimes$ и $\oplus$), то он может нарушаться, и при этом довольно значительно. Пусть, например, $u=20000.000$, $v=-6.0000000$ и $w=6.0000003$, тогда

$$(u\otimes v)\oplus(u\otimes w)=120000.00\oplus 120000.01=.010000000,$$
$$u\otimes(v\oplus w)=20000.000\otimes .00000030000000=.0060000000,$$

так что

$$u\otimes(v\oplus w)\ne(u\otimes v)\oplus(u\otimes w). \tag{15}$$

Аналогично, нетрудно указать примеры, когда $2(u^2\oplus v^2)<(u\oplus v)^2$; запрограммированное вычисление среднего квадратичного отклонения для ряда наблюдений по формуле

$$\sigma=\frac{1}{n}\sqrt{n\sum_{1\le k\le n}x_k^2-\left(\sum_{1\le k\le n}x_k\right)^2}$$

может привести к извлечению квадратного корня из отрицательного числа!

Даже если алгебраические законы выполняются не вполне строго, мы можем использовать наши методы для определения того, с какой степенью точности выполняется закон. Из определения $\mathrm{round}(x,p)$ следует, что

$$\mathrm{round}(x,p)=x(1+\sigma_p(x)), \tag{16}$$

где

$$|\sigma_p(x)|\le\frac{1}{2}b^{1-p}. \tag{17}$$

Следовательно, мы всегда можем записать

$$a\oplus b=(a+b)(1+\delta_p(a+b)), \tag{18}$$
$$a\ominus b=(a-b)(1+\delta_p(a-b)), \tag{19}$$
$$a\otimes b=(a\times a)(1+\delta_p(a\times b)), \tag{20}$$
$$a\oslash b=(a/b)(1+\delta_p(a/b)). \tag{21}$$

Здесь довольно простым способом можно оценить относительную ошибку нормализованных вычислений в системе с плавающей точкой. Формулы (18)–(21) служат главным инструментом для оценки ошибок в арифметике нормализованных чисел с плавающей точкой.

В качестве примера типичной процедуры оценки ошибки рассмотрим закон ассоциативности умножения. Как показывает упр. 3, $(u\times v)\otimes w$, вообще говоря, не равно $u\otimes(v\otimes w)$; но ситуация в данном случае намного лучше, чем в случае закона ассоциативности сложения (1) и закона дистрибутивности (15). В самом деле, ввиду (17) и (20) имеем

$$(u\otimes v)\otimes w=((uv)(1+\delta_1))\otimes w=uvw(1+\delta_1)(1+\delta_2),$$
$$u\otimes(v\otimes w)=u\otimes((vw)(1+\delta_3))=uvw(1+\delta_3)(1+\delta_4)$$

для некоторых δ_1, δ_2, δ_3, δ_4 при условии, что не происходит переполнения или исчезновения показателя, причем $|\delta_j| \le \frac{1}{2}b^{1-p}$ для каждого j.

Следовательно,

$$\frac{(u \otimes v) \otimes w}{u \otimes (v \otimes w)} = \frac{(1+\delta_1)(1+\delta_2)}{(1+\delta_3)(1+\delta_4)} = 1+\delta,$$

где

$$|\delta| \le 2b^{1-p} / \left(1 - \frac{1}{2}b^{1-p}\right)^2. \tag{22}$$

Тем самым мы установили, что $(u \otimes v) \otimes w$ ***приблизительно равно*** $u \otimes (v \otimes w)$, за исключением тех случаев, когда происходит исчезновение или переполнение показателя. Эта интуитивная идея "приблизительного равенства" заслуживает более подробного изучения; можно ли разумным образом дать более точную формулировку этого утверждения?

Программист, использующий арифметические операции в системе с плавающей точкой, почти никогда не испытывает желания проверить, не выполняется ли равенство $u = v$ (или по крайней мере он едва ли когда-нибудь пытается это сделать), так как равенство является предельно маловероятным событием. Например, если используется рекуррентное соотношение

$$x_{n+1} = f(x_n),$$

о котором теория, взятая из какой-то книжки, утверждает, что x_n стремится к некоторому пределу при $n \to \infty$, то, как правило, было бы ошибкой продолжать вычисления, пока для некоторого n не осуществится равенство $x_{n+1} = x_n$, так как последовательность x_n может ввиду округления промежуточных результатов оказаться периодической с большим периодом. Разумно продолжать вычисления лишь до тех пор, пока для некоторого подходящим образом выбранного δ не станет справедливо неравенство $|x_{n+1} - x_n| < \delta$; но так как мы не знаем заранее порядка величины x_n, еще более правильно дожидаться появления неравенства

$$|x_{n+1} - x_n| \le \varepsilon |x_n|; \tag{23}$$

число ε гораздо легче выбрать заранее. Соотношение (23)—это другой способ выражения того факта, что числа x_{n+1} и x_n приблизительно равны; и наше обсуждение показывает, что при рассмотрении вычислений над числами с плавающей точкой соотношение "приблизительного равенства" было бы более полезно, чем традиционное соотношение равенства, если только нам удастся определить первое соотношение надлежащим образом.

Другими словами, тот факт, что строгое равенство величин в системе с плавающей точкой играет очень небольшую роль, приводит к необходимости ввести новую операцию ***сравнения величин с плавающей точкой,*** предназначенную для облегчения оценок относительных значений двух таких величин. Представляются пригодными следующие определения для чисел с плавающей точкой $u = (e_u, f_u)$ и $v = (e_v, f_v)$ по основанию b с избытком q:

$$u \prec v(\varepsilon) \text{ тогда и только тогда, когда } v - u > \varepsilon \max(b^{e_u - q}, b^{e_v - q}); \tag{24}$$

$$u \sim v(\varepsilon) \text{ тогда и только тогда, когда } |v - u| \le \varepsilon \max(b^{e_u - q}, b^{e_v - q}); \tag{25}$$

$$u \succ v(\varepsilon) \text{ тогда и только тогда, когда } u - v > \varepsilon \max(b^{e_u - q}, b^{e_v - q}); \tag{26}$$

$$u \approx v(\varepsilon) \text{ тогда и только тогда, когда } |v - u| \le \varepsilon \min(b^{e_u - q}, b^{e_v - q}). \tag{27}$$

Согласно этим определениям, для любой данной пары значений u, v может выполняться в точности одно из соотношений $u \prec v$ ("определенно меньше"), $u \sim v$ ("приблизительно равно") или $u \succ v$ ("определенно больше"). Соотношение $u \approx v$—несколько более сильное, нежели $u \sim v$, и его можно читать так: "u по существу равно v". Все эти соотношения задаются посредством положительного числа ε, измеряющего степень рассматриваемого приближения.

Одним из способов истолкования приведенных определений состоит в том, чтобы любому числу u с плавающей точкой поставить в соответствие множество $S(u)\{\, x \mid |x - u| \le \varepsilon b^{e_u - q} \,\}$; множество $S(u)$ представляет собой совокупность вещественных чисел, расположенных вблизи u, и определено при помощи показателя u. В терминах этих множеств соотношение $u \prec v$ выполняется тогда и только тогда, когда $S(u) < v$ и $u < S(v)$; $u \sim v$ тогда и только тогда, когда $u \in S(v)$ или $v \in S(u)$; $u \succ v$ тогда и только тогда, когда $u > S(v)$ и $S(u) > v$; $u \approx v$ тогда и только тогда, когда $u \in S(v)$ и $v \in S(u)$. (Здесь мы предполагаем, что параметр ε, измеряющий степень приближения, фиксирован; в более подробной записи можно было отразить и зависимость $S(u)$ от ε.)

Вот некоторые следствия из приведенных определений:

$$\text{если } u \prec v(\varepsilon), \text{ то } v \succ u(\varepsilon); \tag{28}$$
$$\text{если } u \approx v(\varepsilon), \text{ то } u \sim v(\varepsilon); \tag{29}$$
$$u \approx u(\varepsilon); \tag{30}$$
$$\text{если } u \prec v(\varepsilon), \text{ то } u < v; \tag{31}$$
$$\text{если } u \prec v(\varepsilon_1) \text{ и } \varepsilon_1 \geq \varepsilon_2, \text{ то } u \prec v(\varepsilon_2); \tag{32}$$
$$\text{если } u \sim v(\varepsilon_1) \text{ и } \varepsilon_1 \leq \varepsilon_2, \text{ то } u \sim v(\varepsilon_2); \tag{33}$$
$$\text{если } u \approx v(\varepsilon_1) \text{ и } \varepsilon_1 \leq \varepsilon_2, \text{ то } u \approx v(\varepsilon_2); \tag{34}$$
$$\text{если } u \prec v(\varepsilon_1) \text{ и } v \prec w(\varepsilon_2), \text{ то } u \prec w(\varepsilon_1 + \varepsilon_2); \tag{35}$$
$$\text{если } u \approx v(\varepsilon_1) \text{ и } v \approx w(\varepsilon_2), \text{ то } u \sim w(\varepsilon_1 + \varepsilon_2). \tag{36}$$

Далее, можно без труда доказать, что

$$|u - v| \leq \varepsilon|u| \text{ и } |u - v| \leq \varepsilon|v| \text{ влекут } u \approx v(\varepsilon), \tag{37}$$
$$|u - v| \leq \varepsilon|u| \text{ или } |u - v| \leq \varepsilon|v| \text{ влекут } u \sim v(\varepsilon) \tag{38}$$

и, обратно, для *нормализованных* чисел u и v при $\varepsilon < 1$

$$u \approx v(\varepsilon) \text{ влечет } |u - v| \leq b\varepsilon|u| \text{ и } |u - v| \leq b\varepsilon|v|; \tag{39}$$
$$u \sim v(\varepsilon) \text{ влечет } |u - v| \leq b\varepsilon|u| \text{ или } |u - v| \leq b\varepsilon|v|. \tag{40}$$

В качестве примера всех этих соотношений имеем [как и в (22)]

$$|(u \otimes v) \otimes w - u \otimes (v \otimes w)| = |u \otimes (v \otimes w)| \cdot \left| \frac{(1+\delta_1)(1+\delta_2)}{(1+\delta_3)(1+\delta_4)} - 1 \right| \leq$$
$$\leq \frac{2b^{1-p}}{\left(1 - \frac{1}{2}b^{1-p}\right)^2} |u \otimes (v \otimes w)|,$$

и то же самое неравенство выполняется, если поменять местами $(u \otimes v) \otimes w$ и $u \otimes (v \otimes w)$. Следовательно, ввиду (27), справедливо соотношение

$$(u \otimes v) \otimes w \approx u \otimes (v \otimes w)(\varepsilon) \tag{41}$$

для $\varepsilon \geq 2b^{1-p}/(1 - b^{1-p})^2$. Например, при $b = 10$ и $p = 8$ можно взять $\varepsilon = 0.00000021$. Таким образом, у нас имеется довольно хороший "закон ассоциативности".

Соотношения $\prec$, $\sim$, $\succ$ и $\approx$ полезны для численных алгоритмов, и поэтому разумна идея обеспечения ЭВМ программами сравнения чисел с плавающей точкой наряду с программами выполнения над ними арифметических действий.

Теперь вновь переключим наше внимание на вопрос о нахождении *точных* соотношений, которым удовлетворяют операции над величинами с плавающей точкой. Интересно отметить, что сложение и вычитание таких величин не полностью выпадают из поля зрения аксиоматики, так как они удовлетворяют нетривиальным тождествам, формулируемым в теоремах A и B.

Теорема A. *Пусть u и v—нормализованные числа с плавающей точкой. Тогда*

$$((u \oplus v) \ominus u) + ((u \oplus v) \ominus ((u \oplus v) \ominus u)) = u \oplus v, \tag{42}$$

при условии, что не происходит переполнения или исчезновения показателя.

Замечание. Это довольно громоздкое тождество можно переписать в следующем более простом виде. Положим

$$\begin{aligned} u' &= (u \oplus v) \ominus v, \qquad & v' &= (u \oplus v) \ominus u; \\ u'' &= (u \oplus v) \ominus v', \qquad & v'' &= (u \oplus v) \ominus u'. \end{aligned} \tag{43}$$

Интуитивно ясно, что u' и u'' должны быть приближениями к u, а v' и v''—приближениями к v. Теорема A утверждает, что

$$u \oplus v = u' + v'' = u'' + v'. \tag{44}$$

Это более сильное утверждение, нежели тождество

$$u \oplus v = u' \oplus v'' = u'' \oplus v', \tag{45}$$

являющееся еще одним следствием теоремы A (см. упр. 12).

Доказательство. Ввиду симметричности наших предположений достаточно установить справедливость равенств (44) при условии $u \geq |v|$. В последующем доказательстве удобно будет использовать сокращения

$$d = e_u - e_v \geq 0, \qquad w = u \oplus v \geq 0 \tag{46}$$

и работать с целыми числами вместо дробей; если малая латинская буква x обозначает нормализованную величину (e_x, f_x) с плавающей точкой, то соответствующая заглавная буква X будет обозначать число $b^{p+e_x-e_v} f_x$. В частности, $U = b^{p+d} f_u$, $V = b^p f_v$; эти величины, равно как и U', V', U'', V'' и W, являются целыми числами, десятичные точки которых выровнены таким образом, что (44) эквивалентно

$$W = U' + V'' = U'' + V'. \tag{47}$$

Теперь доказательство сводится к довольно скучному рассмотрению ряда частных случаев.

Случай 1: $e_w = e_u$. (См. рис. 4 (i).) Здесь $U + V = W + R$, где

$$R \equiv V \pmod{b^d}, \qquad -\frac{1}{2}b^d \leq R < \frac{1}{2}b^d.$$

Имеем $U' = \operatorname{round}(W - V, p) = \operatorname{round}(U - R, p)$. Далее возможны два подслучая. *Случай (1a):* $R = -\frac{1}{2}b^d$. Тогда $U' = U + b^d$,

$U = \ u u u u u u u u 0 0$	$U = \ u u u u u u u u 0 0$	$U = \ u u u u u u u u 0 0$
$V = \ \ \ v v v v v v v v$	$V = \ \ \ v v v v v v v v$	$V = \ \ \ v v v v v v v v$
$W = w w w w w w w w 0 0$	$W = w w w w w w w 0 0 0$	$W = w w w w w w w w w 0$
(i)	(ii)	(iii)

Picture: Рис. 4. Возможные случаи выравнивания позиционной точки при сложении.

$V' = V - R$, $U'' = U$, $V'' = V - R - b^d = V - \frac{1}{2}b^d$. *Случай (1b):* $R \neq -\frac{1}{2}b^d$. Тогда $U' = U$, $V' = V - R$, $U'' = U$, $V'' = V - R$.

Случай 2: $e_w = e_u + 1$. (См. рис. 4 (ii).) Ясно, что $V > 0$ и $d \leq p$. Имеем $U + V = W + R$, где

$$R \equiv V + b^d U_0 \pmod{b^{d+1}}, \qquad -\frac{1}{2} \leq R < \frac{1}{2}b^{d+1},$$

а U_0—наименее значимая цифра f_u. Снова рассмотрим подслучаи. *Случай (2a):* $U - R \geq b^{d+p} - \frac{1}{2}b^d$. Так как $U - R \leq b^{d+p} - b^d + \frac{1}{2}b^{d+1}$, должно выполняться равенство $U' = b^{d+p} = U - R + Q$, где

$$Q \equiv V \pmod{b^{d+1}}, \qquad -\frac{1}{2}b^{d+1} < Q \leq \frac{1}{2}b^{d+1}.$$

Следовательно, $V'' = V - Q$. *Случай (2b):* $b^{d+p} - \frac{1}{2}b^d > U - R \geq b^{d+p-1} - \frac{1}{2}b^{d-1}$. Тогда $U' = U - R + Q$, где

$$Q \equiv V \pmod{b^d}, \qquad -\frac{1}{2}b^d < Q \leq \frac{1}{2}b^d,$$

и $V'' = V - Q$. *Случай (2c):* $b^{d+p-1} - \frac{1}{2}b^{d-1} > U - R$. Этот случай невозможен. Это очевидно, когда $d = 0$. А если $d > 0$, то $R > 0$, так что $U + V > W$ и $U - R \geq W - V - R + 1 > b^{d+p} - (b^p - 1) - \frac{1}{2}b^{d+1} + 1 \geq b^{d+p-1}$, и мы приходим к противоречию. Чтобы, наконец, завершить анализ случая 2, мы должны вычислить $V' = \operatorname{round}(V - R, p)$. Здесь $V - R$ содержит не более $p + 1$ разрядов, причем d наименее значимых цифр равны нулю, так что если $d \neq 0$, то $V' = V - R$, $U'' = U$. Если же $d = 0$, то обязательно $V' = V - R$, за исключением того случая, когда $V' = b^{p+d}$, и в этом последнем случае имеет место та необычная ситуация, когда W принимает свое максимальное значение $2b^{p+d}$; здесь $U'' = b^{p+d}$ и $b > 3$.

Случай 3: $e_w < e_u$. (См. рис. 4 (iii).) Здесь $V < 0$. *Случай (3a):* $d \leq 1$. Тогда $U + V = W$, так что $U' = U'' = U$, $V' = V'' = V$. *Случай (3b):* $d > 1$. Тогда $e_w = e_u - 1$ и $U + V = W + R$, где

$$R \equiv V \pmod{b^{d-1}}, \quad -\frac{1}{2}b^{d-1} \leq R < \frac{1}{2}b^{d-1}.$$

Этот случай аналогичен случаю 1, но проще, ввиду того что интервал изменения R меньше. Имеем $U' = U$, $V' = V - R$, $U'' = U$, $V'' = V - R$. ∎

Теорема A выявляет некое свойство регулярности операции сложения в системе плавающей точкой, но она не представляется особенно полезным результатом. Следующая теорема гораздо более существенна.

Теорема B. *В предположениях теоремы A и при условии (43) справедливо тождество*

$$u + v = (u \oplus v) + ((u \ominus u') \oplus (v \ominus v'')). \tag{48}$$

Доказательство. Рассматривая каждый из случаев, возникших при доказательстве теоремы A, мы неизменно обнаруживаем, что

$$\begin{aligned} u \ominus u' = u - u', \quad v \ominus v' = v - v', \\ u \ominus u'' = u - u'', \quad v \ominus v'' = v - v'', \\ ((u \ominus u') \oplus (v \ominus v'')) = ((u - u') + (v - v'')) = \\ = ((u - u'') + (v - v')) = \\ = ((u \ominus u'') \oplus (v \ominus v')), \end{aligned}$$

поскольку каждую из этих величин можно точно выразить как p-разрядное число с плавающей точкой, без всякого округления. Например, в случае 2 имеем $U - U' = R - Q \equiv O \pmod{b^d}$, и во всяком случае $|R| < b^p$. Комбинируя теперь вышеприведенные тождества с теоремой A, получаем доказываемое утверждение. ∎

Теорема B дает нам *явную формулу для разности* между $u + v$ и $u \oplus v$ в терминах величин, которые допускают прямое вычисление при помощи операций однократной точности над величинами с плавающей точкой. Поэтому часто можно увеличить точность вычисления однократной точности в системе с плавающей точкой, накапливая поправочные члены $(u \ominus u') \oplus (v \ominus v'')$. Читатель, который старательно проследил за всеми деталями доказательства теорем A и B, поймет, сколь значительные упрощения дало простое правило

$$u \oplus v = \operatorname{round}(u + v, p).$$

Если бы наша программа сложения чисел с плавающей точкой не обеспечивала этого результата всегда, то, сколь бы редки ни были исключения, все доказательство стало бы несравненно более сложным, а возможно, и просто не проходило бы.

Теорема B была бы неверна, если бы мы использовали арифметику с урезанием числа вместо округления, т. е. если бы в соотношениях **(11)–(15)** мы подставили[22] $\operatorname{trunc}(x, p)$ вместо $\operatorname{round}(x, p)$; функция $\operatorname{trunc}(x, p)$ похожа на функцию $\operatorname{round}(x, p)$ при $x > 0$ с тем лишь отличием, что величина $\frac{1}{2}$, фигурирующая в **(9)**, заменяется нулем.

Теоремой B не охватывались бы тогда случаи типа, например,

$$(20, +.10000001) \oplus (10, -.10000001) = (20, +.10000000),$$

когда разность между $u + v$ и $u \oplus v$ нельзя было бы точно выразить как число с плавающей точкой. Если бы урезание производилось каким-либо иным способом, то при использовании такого урезания в средней части алгоритма 4.2.1A без ограничений независимо от знака чисел могло бы случиться, что теоремы A и B остались верными, но получающаяся при этом операция $\oplus$ оказалась бы много менее доступной для математического анализа.

Многие думают, что, поскольку "плавающая арифметика" неточна по самой своей природе, не будет никакой беды в том, чтобы в некоторых довольно редких случаях выполнять ее операции чуть менее точно, если это окажется удобным. Такая политика сберегает несколько центов при проектировании ЭВМ или небольшой процент общего времени работы подпрограммы. Однако проведенное нами выше исследование показывает, что такой подход ошибочен. Даже при условии, что скорость подпрограммы FADD программы 4.2.1A, если бы мы допустили возможность неверного округления в небольшом числе случаев, возросла бы, скажем, на пять процентов, все равно гораздо лучше оставить ее такой, как она есть. И дело здесь не в "погоне за битами" и не в том, чтобы в средней программе получать фантастически хорошие результаты; на карту поставлено нечто более важное и фундаментальное: *числовые подпрограммы должны давать результаты, которые, насколько это возможно удовлетворяют простым полезным математическим законам.* Ключевая формула $u \oplus v = \operatorname{round}(u + v, p)$, например, выражает некое свойство "регулярности", и этим решается вопрос, стоит проводить математический анализ вычислительных алгоритмов или не стоит. Не располагая какими-либо лежащими в основе свойствами симметрии, доказывать интересные результаты было бы крайне неудобно. Быть довольным инструментом, которым работаешь,—это, конечно, существенное условие успешной работы.

---

[22] Ниже trunc—от английского truncation (урезание).—*Прим. ред.*

B. Арифметические действия над ненормализованными числами с плавающей точкой. К стратегии нормализации всех чисел с плавающей точкой можно относиться двояко: либо благосклонно воспринимать ее как попытку получить минимальные погрешности, достижимые для данной степени точности, либо рассматривать ее как потенциально опасную линию поведения в том смысле, что при этом имеется тенденция выдавать результаты за более точные, чем они есть на самом деле. Когда мы, нормализуя результат операции $(1, +.31428571) \ominus (1, +.31415927)$, получаем $(-2, +.12644000)$, мы теряем информацию о максимальной степени неточности последней величины. Такая информация сохранилась бы, если бы мы оставили ответ в виде $(1, +.00012644)$.

Входные данные к задаче часто неизвестны с той точностью какая может допускаться представлением с плавающей точкой Например, значения числа Авогадро и постоянной Планка с восемью значащими цифрами неизвестны, и было бы более удобно обозначать их

$$(27, +.00060225) \text{ и } (-23, +.00010545)$$

соответственно, а не $(24, +.60225000)$ и $(-26, +.10545000)$. Было бы весьма приятно, если бы мы могли задавать наши входные данные для каждой задачи в ненормализованной форме, которая бы отражала степень принятой точности, и если бы в выходных данных имелась информация о том, какова точность ответа. К несчастью, это ужасно трудная проблема, хотя использование ненормализованной арифметики и может помочь нам получить некоторые указания такого рода. Например, мы можем сказать с большой степенью уверенности, что произведение числа Авогадро на постоянную Планка равно $(0, +.00063507)$, а их сумма равна $(27, +.00060225)$. (Назначение этого примера не в том, чтобы навести на мысль, что можно приписать какой-либо важный физический смысл сумме или произведению этих фундаментальных постоянных; суть в том, что можно сохранить немного информации о точности результата вычислений над неточными величинами, когда исходные операнды не зависят один от другого.)

Правила ненормализованной арифметики просты и состоят в следующем: пусть l_u—количество нулей, стоящих в начале дробной части величины $u = (e_u, f_u)$, так что l_u есть наибольшее целое число $\le p$, для которого $|f_u| < b^{-l_u}$. Тогда сложение и вычитание выполняются в точности так же, как и в алгоритме 4.2.1A, за тем исключением, что все сдвиги опускаются. Умножение и деление выполняются в точности так же, как в алгоритме 4.2.1M, за тем исключением, что ответ сдвинут вправо или влево, так что дробная часть ответа будет начинаться в точности с $\max(l_u, l_v)$ нулей. По существу те же правила использовались и для традиционных вычислений вручную.

Таким образом, для ненормализованных вычислений мы имеем

$$e_{u\oplus v}, e_{u\ominus v} = \max(e_u, e_v) + (0 \text{ или } 1), \tag{49}$$

$$e_{u\otimes v} = e_u + e_v - q - \min(l_u, l_v) - (0 \text{ или } 1), \tag{50}$$

$$e_{u\oslash v} = e_u - e_v + q - l_u + l_v + \max(l_u, l_v) + (0 \text{ или } 1). \tag{51}$$

Когда итогом вычислений является нуль, то в качестве результата выдается ненормализованный нуль (часто называемый "величиной порядка нуль"); это означает, что в действительности результат может быть и не равен нулю, но мы просто не знаем ни одной его значащей цифры!

При использовании арифметических операций над ненормализованными числами с плавающей точкой формулы (18)–(21), использовавшиеся нами для анализа ошибок, перестают быть верными. Введем теперь новую величину

$$\delta_u = \frac{1}{2} b^{e_u - q - p} \qquad \text{для } u = (e_u, f_u). \tag{52}$$

Эта величина зависит от представления числа u, а не от его значения $b^{e_u - q} f_u$. Тогда вместо (18)–(21) мы имеем неравенства

$$|u \oplus v - (u + v)| \le \delta_{u\oplus v}, \tag{53}$$

$$|u \ominus v - (u - v)| \le \delta_{u\ominus v}, \tag{54}$$

$$|u \otimes v - (u \times v)| \le \delta_{u\otimes v}, \tag{55}$$

$$|u \oslash v - (u/v)| \le \delta_{u\oslash v}. \tag{56}$$

Эти неравенства являются простым следствием правила округления, и они применимы как к нормализованным, так и к ненормализованным величинам; основное различие между двумя типами анализа ошибок состоит в определении показателя результата каждой операции (соотношения (49)–(51)).

Отношения $\prec$, $\sim$, $\succ$ и $\approx$, определенные нами выше в (24)–(27), сохраняют смысл и для ненормализованных чисел. В качестве примера использования этих отношений докажем приближенный закон ассоциативности для сложения ненормализованных величие (аналогичный закону (41)): для подходящим образом выбранного ε

$$(u \oplus v) \oplus w \approx u \oplus (v \oplus w)\,(\varepsilon). \tag{57}$$

Имеем

$$\begin{aligned}|(u \oplus v) \oplus w - (u + v + w)| &\le |(u \oplus v) \oplus w - ((u \oplus v) + w)| + |u \oplus v - (u + v)| \le \\ &\le \delta_{(u\oplus v)\oplus w} + \delta_{u\oplus v} \le \\ &\le 2\delta_{(u\oplus v)\oplus w}.\end{aligned}$$

Аналогичная формула справедлива и для $|u \oplus (v \oplus w) - (u + v + w)|$. Но поскольку $e_{(u\oplus v)\oplus w} = \max(e_u, e_v, e_w)+$ $(0, 1$ или $2)$, то $\delta_{(u\oplus v)\oplus w} \le b^2\delta_{u\oplus(v\oplus w)}$. Следовательно, (57) верно при $\varepsilon \ge 2b^{2-p}$; с точки зрения закона ассоциативности сложение ненормализованных величин не столь ошибочно, как сложение нормализованных.

Следует подчеркнуть, что арифметика ненормализованных величин ни в коей мере не может служить панацеей; имеются примеры, где указываемая такой арифметикой точность больше действительной (например, сложение большого числа малых приблизительно одинаковых величин или нахождение n-й степени числа для большого n), и существует еще больше примеров, когда в этой арифметике дается невысокая точность, в то время как в арифметике над нормализованными величинами в действительности получались бы гораздо более точные результаты.

Есть важная причина, по которой ни один прямой метод анализа ошибок "по операции зараз" не надежен, а именно тот факт, что операнды обычно не являются независимыми. Это означает, что ошибки имеют тенденцию компенсировать или усиливать друг друга необычным образом. Предположим, например, что x приблизительно равно $1/2$, и предположим, что у этого значения есть приближение $y = x + \delta$ с абсолютной ошибкой δ. Если мы хотим вычислить произведение $x(1 - x)$, то мы можем найти $y(1-y)$; если $x = \frac{1}{2} + \varepsilon$, то мы получим $y(1-y) = x(1-x) - 2\varepsilon\delta - \delta^2$. Благодаря множителю 2ε ошибка уменьшилась! Это только один из случаев, когда умножение приближенных значений может привести к весьма точному результату, если операнды не являются независимыми. Более очевидный пример—это вычисление $x \ominus x$, которое может быть выполнено с абсолютной точностью независимо от того, насколько плохим было приближение x, с которого начинали. Дополнительная информация, которую дает нам арифметика ненормализованных величин, часто может быть более важна, нежели та информация, которая теряется при проведении обширных вычислений в такой арифметике, но (как обычно) необходимо соблюдать осторожность при ее использовании. Примеры правильного использования арифметики ненормализованных величин обсуждались Р. Эшенхёрстом и Н. Метрополисом в сборнике "Computers and Computing" [AMM Slaught Memorial Papers, 10 (February, 1965), 47–59] и Р. Эшенхёрстом в сборнике "Error in digital computation" [vol. II. ed. by L. B. Rall, New York, Wiley, 1965, 3–37]. Надлежащие методы вычисления стандартных математических функций с представлением как входных, так и выходных данных в ненормализованной форме были даны Р. Эшенхёрстом в *JACM* [**11** (1964), 168–187].

Другой подход к проблеме оценки погрешности связан с так называемым "интервальным" методом, в котором вычисления проводятся с учетом верхней и нижней оценок для каждого числа. Так, например, если известно, что $u_0 \le u \le u_1$ и $v_0 \le v \le v_1$, то $u_0+v_0 \le u+v \le u_1+v_1$, $u_0-v_1 \le u-v \le u_1-v_0$ и (в предположении, что u_0, v_0 положительны) $u_0v_0 \le uv \le u_1v_1$, $u_0/v_1 \le u/v \le u_1/v_0$. Аналогичные правила для других случаев получаются непосредственно; разумеется, возникают трудности, когда $v_0 < 0 < v_1$, а мы пытаемся делить на v. Вычисления над u_0, u_1, v_0 и v_1 проводятся с соответствующим округлением (вниз для нижних границ и вверх для верхних). Такие вычисления лишь вдвое больше по объему, чем те же вычисления в обычной арифметике, и так как они обеспечивают точную оценку ошибки, они могут оказаться весьма ценными. Однако ввиду зависимости промежуточных значений друг от друга окончательные оценки часто слишком пессимистичны; имеются также некоторые проблемы, связанные с применением итерационных численных методов. По поводу обсуждения интервального метода и некоторых его модификаций см. статьи Э. Гибба [*CACM,* **4** (1961), 319–320] и Б. Шартра [*JACM,* **13** (1966), 386–403], а также книгу Р. Мура "Interval analysis" [Englewood Cliffs, Prentice Hall, 1966].

С. История и библиография. Первое исследование плавающей арифметики было выполнено Ф. Л. Бауэром и К. Замельзоном [Optimale Rechengenauigkeit bei Rechenanlagen mit gleitendem Komma, *Zeitschrift für angewandte Math. und Physik,* **4** (1953), 312–316]. Следующая публикация появилась лишь пятью годами позже [J. W. Carr III, Error analysis in floating-point arithmetic, *CACM,* **2** (May, 1959), 10–15]. См. также [P. C. Fischer, Proc. ACM 13th Nat. Meeting, Urbana, Illinois, 1958,

paper 39]. В книге Дж. Х. Уилкинсона "Rounding errors in algebraic processes" [Englewood Cliffs, Prentice-Hall, 1963] показано, как применять методы анализа ошибок индивидуальных арифметических операций к анализу ошибок в задачах с большим числом операций; см. также его монографию "The algebraic eigenvalue problem" [Oxford, Clarendon Press, 1965].

Введенные в этом пункте отношения $\prec$, $\sim$, $\succ$, $\approx$ сородственны идеям, провозглашенным А. ван Вейнгаарденом [Numerical analysis as an independent science, *BIT,* **6** (1966), 66–81]. Приведенные выше теоремы A и B навеяны некоторыми близкими результатами Уле Мёллера [*BIT,* **5** (1965), 37–50, 251–255]. См. также [W. Kahan, *CACM,* **8** (1965), 40].

В пользу арифметики ненормализованных чисел с плавающей точкой выступили Ф. Л. Бауэр и К. Замельзон в упомянутой выше статье, и независимо ее использовал Дж. В. Карром из Мичиганского университета (1953 г.). Несколькими годами позже была спроектирована машина MANIAC III со схемной реализацией арифметики обоих типов, см. R. L. Ashenhurst, N. Metropolis, *JACM,* **6** (1959), 415–428; *IEEE Transactions on Electronic Computers,* **EC-12** (1963), 896–901; R. L. Ashenhurst, Proc. Spring Joint Computer Conf., **21** (1962), 195–202. По поводу других ранних обсуждений ненормализованной арифметики см. также H. L. Gray, C. Harrison, Jr., Proc. Eastern Joint Computer Conf., **16** (1959), 244–248, и W. G. Wadey, *JACM,* **7** (1960), 129–139.

Упражнения

(В этих задачах предполагается, если не оговорено противное, что действия выполняются над нормализованными числами с плавающей точкой.)

1. [M18] Докажите, что тождество (7) следует из соотношений (2)–(6).
2. [M20] Используя тождества (2)–(8), докажите, что $(u \oplus x) \oplus (v \oplus y) \geq u \oplus v$, каковы бы ни были $x \geq 0$ и $y \geq 0$.
3. [M20] Найдите восьмиразрядные десятичные числа с плавающей точкой u, v и w, для которых

$$u \otimes (v \otimes w) \neq (u \otimes v) \otimes w,$$

причем ни при одном из этих вычислений не происходит ни переполнения, ни исчезновения показателя.

4. [10] Можно ли найти числа с плавающей точкой u, v и w, для которых при вычислении $u \times (v \times w)$ происходило бы исчезновение показателя, а при вычислении $(u \otimes v) \otimes w$ не происходило?
5. [M20] Выполняется ли равенство $u \oslash v = u \otimes (1 \oslash v)$ для всех чисел с плавающей точкой u и $v \neq 0$, для которых не возникает ни переполнения, ни исчезновения показателя?
6. [M22] Для каждого из следующих двух соотношений выясните, выполняется ли оно тождественно для всех чисел с плавающей точкой u. (a) $0 \ominus (0 \ominus u) = u$; (b) $1 \oslash (1 \oslash u) = u$.
7. [M20] Докажите, что для $\delta_p(x)$, определенного соотношением (16), справедливо неравенство (17).

>8. [20] Пусть $\varepsilon = 0.0001$; какое из соотношений $u \prec v(\varepsilon)$, $u \sim v(\varepsilon)$, $u \succ v(\varepsilon)$, $u \approx v(\varepsilon)$ выполняется для следующих пар восьмиразрядных десятичных чисел с плавающей точкой с избытком 0?

   a) $u = (1, +.31415927)$, $v = (1, +.31416000)$;
   b) $u = (0, +.99997000)$, $v = (1, +.10000039)$;
   c) $u = (24, +.60225200)$, $v = (27, +.00060225)$;
   d) $u = (24, +.60225200)$, $v = (31, +.00000006)$;
   e) $u = (24, +.60225200)$, $v = (32, +.00000000)$.

9. [M22] Докажите утверждение (36) и объясните, почему заключение нельзя усилить до $u \approx w(\varepsilon_1 + \varepsilon_2)$.

>10. [M25] (У. Кахан.) На некоторой ЭВМ выполнение арифметических действий над числами с плавающей точкой проводится без точного округления, и фактически программа умножения для этой ЭВМ игнорирует последние p разрядов $2p$-разрядного произведения $f_u f_v$. (Таким образом; если $f_u f_v < 1/b$, то из-за последующей нормализации наименее значимая цифра всегда оказывается нулем.) Покажите, что это приводит к утрате монотонности умножения, т. е. что существуют такие положительные нормализованные числа с плавающей точкой u, v, w, что $u < v$, но $u \otimes w > v \otimes w$.

>11. [M28] Вместо того чтобы использовать для дробных частей чисел с плавающей точкой прямой код, мы могли бы следующим образом воспользоваться дополнительным кодом (см. § 4.1). Дробная часть f положительного числа находится, как и раньше, в интервале $(0.100\ldots0)_2 = 1/2 \leq f \leq 1 - 2^{-p} = (0.111\ldots1)_2$, но дробная часть f *отрицательного* числа лежит в интервале $(1.000\ldots0)_2 = -1 \leq f \leq -1/2 - 2^{-p} = (1.011\ldots1)_2$. Сложение и вычитание можно выполнять при помощи такого непосредственного обобщения алгоритма 4.2.1A: обеспечивая достаточную

точность вычислений, мы получаем верное значение суммы или разности, потом нормализуем дробь, так чтобы ее первые p разрядов имели надлежащий вид, а после этого "округляем" результат, добавляя единицу в $(p+1)$-й разряд, и затем отбрасываем все разряды, кроме первых p битов, производя в случае переполнения при округлении денормализацию результата. Например, разность $(2, 0.11111111) \ominus (6, 0.10000000)$ была бы вычислена сначала в виде $(6, 1.100011111111)$, нормализована к виду $(5, 1.00011111111)$ и затем округлена до $(5, 1.00100000)$. Взяв те же числа в противоположном порядке, мы получили бы

$$(6, 0.10000000) \ominus (2, 0.11111111) = (5, 0.111000000);$$

это предыдущий ответ, взятый с противоположным знаком, так что соотношение (7) выполняется для данного случая.

Найдите два числа u и v, представленные в дополнительном двоичном коде, для которых равенство (7) *не* выполняется и для которых в ходе вычислений не происходит ни переполнения, ни исчезновения показателя.

12. [M15] Почему (45) следует из (44)?

>13. [M25]Некоторые языки программирования (и даже некоторые компьютеры) используют только арифметику над величинами с плавающей точкой и не имеют средств для точных вычислений с целыми числами. Если требуется выполнять операции над целыми числами, мы можем, конечно, представить их в виде чисел с плавающей точкой, и если операции арифметики над числами с плавающей точкой удовлетворяют основным определениям (11)–(14) этого пункта, то, как мы знаем, *все* эти операции оказываются точными, при условии что операнды и ответ допускают точное представление в p-разрядной сетке. Следовательно, пока мы уверены, что числа не слишком велики, мы можем складывать, вычитать или умножать целые числа, не опасаясь неточности, связанной с ошибками округления.

Но предположим, что программист хочет определить, является ли m точным кратным n, где m и $n \neq 0$—целые числа. Предположим далее, что в нашем распоряжении, как и в упр. 4.2.1-15, есть подпрограмма, которая вычисляет $\text{round}(u \bmod 1, p) = u \pmod 1$ для любого числа u с плавающей точкой. Один из хороших способов определить, является ли m кратным n, мог бы состоять в том, чтобы проверить при помощи упомянутой подпрограммы, верно ли равенство $((m \oslash) \pmod 1) = 0$. Не исключено, однако, что ошибки округления в вычислениях над величинами с плавающей точкой сделают эту проверку недостоверной.

Найдите соответствующие ограничения на интервал изменения целых чисел $n \neq 0$ и m, при которых m будет кратным n в том и только том случае, когда $(m \oslash n) \pmod 1 = 0$. Другими словами, покажите, что если m и n не слишком велики, то наша проверка пригодна.

14. [M27] Найдите подходящее значение ε, при котором $(u \otimes v) \otimes w \approx u \otimes (v \otimes w)\ (\varepsilon)$ в случае, когда используется *ненормализованное* умножение. (Это—обобщение соотношения (41), поскольку ненормализованное умножение ничем не отличается от нормализованного, если входные данные u, v и w нормализованы.)
15. [M24] (X. Бьёрк.) Всегда ли вычисленная средняя точка интервала лежит между его концевыми точками? (Иными словами, следует ли из неравенства $u \leq v$ неравенство $u \leq (u \oplus v) \otimes 2 \leq v$?)
16. [BM23] Предположим, что u и v—вещественные числа, независимо и равномерно распределенные в интервалах $0 < u_0 - \delta \leq u < u_0 + \delta$ и $0 < v_0 - \varepsilon \leq v \leq v_0 + \varepsilon$. (a) Каково среднее значение произведения uv? (b) Каково среднее значение частного u/v? [Эти вопросы имеют отношение к вопросу о выборе правильного способа округлять результаты операций умножения и деления.]
17. [28] Напишите MIX-подпрограмму FCMP, которая сравнивает между собой числа u и v в форме с плавающей точкой, находящиеся соответственно в поле ACC и в регистре A, и устанавливает индикатор сравнения в состояния "меньше", "равно" или "больше" в соответствии с тем, будет ли $u \prec v$, $u \sim v$ или $u \succ v (\varepsilon)$; при этом ε хранится в поле EPSILON как неотрицательная величина в форме с плавающей точкой, причем предполагается, что точка расположена слева от слова.
18. [M40]. Существует ли в арифметике ненормализованных величин подходящее число ε, такое, что

$$u \otimes (v \otimes w) \approx (u \otimes v) \otimes (u \otimes w)\ (\varepsilon)?$$

4.2.3. *Вычисления с двойной точностью

До сих пор мы говорили об арифметике чисел с плавающей точкой "однократной точности", что по существу означает, что представленные в форме с плавающей точкой величины, с которыми мы работали, могли храниться в одном машинном слове. Если арифметика однократной точности не обеспечивает достаточной для наших потребностей точности, то точность можно увеличить

при помощи средств программистского характера, используя для представления каждого числа два или больше слов памяти. Хотя общую проблему вычислений повышенной точности мы обсуждаем в § 4.3, имеет смысл отдельно обсудить здесь вопрос о вычислениях двойной точности; к вычислениям двойной точности применимы специальные методы, практически непригодные для случая большей точности; кроме того, вычисления с двойной точностью разумно считать темой, имеющей самостоятельное значение, так как это первый шаг за пределы однократной точности, позволяющий удовлетворительно решать многие задачи, не требующие непомерно высокой точности.

Для выполнения арифметических действий над числами с плавающей точкой двойная точность необходима почти всегда в от-

***** Разрыв текста (пропущены стр. 261-270) *****

это явление было отмечено американским астрономом Саймоном Ньюкомбом [*Amer. J. Math.*, **4** (1881), 39–40], который привел разумные основания в пользу того, что головная цифра d встречается с вероятностью $\log_{10}(1+1/d)$. Тот же самый закон распределения много лет спустя был эмпирически найден Ф. Бенфордом [*Proc. Amer. Philosophical Soc.*, **78** (1938), 551], который не знал о заметке Ньюкомба. Бенфорд решил, что это важный закон природы, и назвал его "законом аномальных чисел". Мы увидим, что этот закон распределения головных цифр является естественным следствием того способа, при помощи которого мы записываем числа в системе с плавающей точкой.

Если мы возьмем произвольное положительное число u, то его головная цифра определяется значением $(\log_{10} u) \bmod 1$. А именно, головная цифра меньше d тогда и только тогда, когда

$$(\log_{10} u) \bmod 1 < \log_{10} d, \tag{1}$$

так как $10f_u = 10^{(\log_{10} u) \bmod 1}$.

Далее, если у нас есть какое-либо "случайное" положительное число U, выбираемое в соответствии с некоторым разумным распределением, типа тех, что встречаются в природе, то можно ожидать, что числа $(\log_{10} U) \bmod 1$ будут равномерно распределены между нулем и единицей или по крайней мере что это будет очень хорошее приближение. (Аналогичным образом мы ожидаем, что величины $U \bmod 1$, $U^2 \bmod 1$, $\sqrt{U+\pi} \bmod 1$ и т. д. также равномерно распределены. Мы уверены, что колесо рулетки беспристрастно по существу по этой же самой причине.) Следовательно, ввиду неравенства (1), головной цифрой будет единица с вероятностью, равной $\log_{10} 2 \approx 30.103\%$, двойка с вероятностью, равной $\log_{10} 3 - \log_{10} 2 \approx 17.609\%$, и вообще если r—произвольное вещественное число, заключенное между 1 и 10, то приблизительно в $\log_{10} r$ всех случаев мы должны иметь неравенство $10f_U \le r$.

Другой способ объяснить этот закон—это сказать, что случайная величина U должна появляться в случайной точке на логарифмической линейке (т. е. что все позиции на логарифмической линейке равновероятны). Действительно, расстояние от левого конца логарифмической линейки до позиции, изображающей число U, пропорционально $(\log_{10} U) \bmod 1$. В случае умножения и деления имеется тесная аналогия между вычислениями, проводимыми при помощи логарифмической линейки, и вычислениями в системе с плавающей точкой.

Тот факт, что головные цифры имеют тенденцию быть небольшими, следует постоянно иметь в виду; именно благодаря этому факту простейшие методы оценки "средней ошибки" годятся для вычислений с плавающей точкой. Относительная ошибка обычно оказывается несколько большей, чем ожидается.

Разумеется, можно справедливо утверждать, что приведенные выше эвристические доводы не доказывают сформулированного закона. Они только указывают правдоподобные причины того, что поведение головных цифр именно таково, каково оно есть на самом деле. Другой подход к анализу головных цифр был предложен Р. С. Пинкэмом и Р. Хэммингом [*Ann Math. Stat.*, **32** (1961), 1223–1230]. Пусть $p(r)$—вероятность того, что $10f_U \le r$, где $1 \le r \le 10$, и f_U—нормализованная дробная часть случайным образом выбранного нормализованного числа U с плавающей точкой. Если говорить о случайных величинах в реальном мире, то мы замечаем, что они измеряются в произвольных единицах, и если бы мы изменили, скажем, определение метра или грамма, то многие бы из фундаментальных физических постоянных имели бы другое значение. Предположим поэтому, что все-все числа во вселенной внезапно оказались умноженными на некоторый постоянный множитель c; наша вселенная случайных величин с плавающей точкой должна после этого преобразования остаться по существу неизменной, так что вероятности $p(r)$ не должны измениться.

Умножение всех чисел на c превращает $(\log_{10} U) \bmod 1$ в $(\log_{10} U + \log_{10} c) \bmod 1$. Настал момент вывести формулы, описывающие искомое распределение; мы можем считать, что $1 \le c \le 10$. По определению

$$p(r) = \text{вероятность}((\log_{10} U) \bmod 1 \le \log_{10} r).$$

Согласно нашему предположению, имеем также

$$\begin{aligned} p(r) &= \text{вероятность}((\log_{10} U + \log_{10} c) \bmod 1 \le \log_{10} r) = \\ &= \begin{cases} \text{вероятность}((\log_{10} U) \bmod 1 \le \log_{10} r - \log_{10} c \\ \text{или } (\log_{10} U) \bmod 1 \ge 1 - \log_{10} c), & \text{если } c \le r, \\ \text{вероятность}(1 - \log_{10} c \le (\log_{10} U) \bmod 1 \le 1 + \log_{10} r - \log_{10} c), & \text{если } c \ge r, \end{cases} \\ &= \begin{cases} p(r/c) + 1 - p(10/c), & \text{если } c \le r, \\ p(10r/c) - p(10/c), & \text{если } c \ge r. \end{cases} \end{aligned} \tag{2}$$

Продолжим теперь функцию $p(r)$ вовне интервала $1 \le r \le 10$, положив $p(10^n r) = p(r) + n$; тогда после замены $10/c$ на d мы можем записать соотношение (2) в виде

$$p(rd) = p(r) + p(d). \tag{3}$$

Если наше предположение об инвариантности распределения относительно умножения на произвольный постоянный множитель верно, то соотношение **(3)** должно выполняться для всех $r > 0$ и $1 \le d \le 10$. Из того что $p(1) = 0$, $p(10) = 1$, следует, что

$$1 = p(10) = p((\sqrt[n]{10})^n) = p(\sqrt[n]{10}) + p((\sqrt[n]{10})^{n-1}) = \\ = \ldots = np(\sqrt[n]{10});$$

отсюда мы заключаем, что для всех положительных целых m и n справедливо равенство $p(10^{m/n}) = m/n$. Если дополнительно потребовать, чтобы распределение p было непрерывным, то мы приходим к равенству $p(r) = \log_{10} r$, а это и есть нужный нам закон.

Хотя это рассуждение, возможно, и убедительнее предыдущих, оно тоже в действительности не выдерживает строгой проверки. Мы предполагаем, что существует некое лежащее в основе рассматриваемого явления распределение чисел $F(u)$, такое, что вероятность того, что данное произвольное число U не превосходит u, равна $F(u)$ и что

$$p(r) = \sum_m (F(10^m r) - F(10^m)), \tag{4}$$

где суммирование проводится по всем значениям $-\infty < m < \infty$. Из нашего рассуждения вытекает, что тогда

$$p(r) = \log_{10} r.$$

Используя те же доводы, мы можем "доказать", что

$$\sum_m (F(b^m r) - F(b^m)) = \log_b r \tag{5}$$

при $1 \le r \le b$ для всякого целого числа $b \ge 2$. Но функции распределения F, которая удовлетворяла бы этому равенству для всех таких b и r, не существует! "Какая-то в державе датской гниль!"

Один из способов выйти из этого затруднения состоит в том, чтобы рассматривать логарифмический закон $p(r) = \log_{10} r$ лишь как очень хорошее *приближение* к истинному распределению. Возможно, что это истинное распределение при расширении Вселенной изменяется, становясь с течением времени все лучшим и лучшим приближением; и если заменить основание 10 произвольным основанием b, наше приближение тем менее точно (в любой данный момент времени), чем больше b. Другой, довольно привлекательный способ решения проблемы, связанный с отказом от традиционного понятия функции распределения, предложен **Р. А. Рэйми [*AMM*, 76 (1969), 342–348].**

Витиеватые рассуждения последнего абзаца, по-видимому, ни в коей мере нельзя признать удовлетворительным объяснением, так что следует весьма положительно отнестись к проводимым ниже вычислениям (где мы придерживаемся строгого математического канона и избегаем интуитивных, но парадоксальных понятии теории вероятностей). Рассмотрим вместо распределения некоего воображаемого множества вещественных чисел распределение старших значащих цифр *положительных целых* чисел. Исследование этой темы чрезвычайно интересно, и не только потому, что оно проливает некоторый свет на распределения вероятностей для данных, представленных в форме с плавающей точкой, но также и потому, что оно служит весьма поучительным примером того, как сочетать методы дискретной математики с методами анализа.

Во всех последующих рассуждениях r будет обозначать фиксированное вещественное число, $1 \le r \le 10$; мы попытаемся дать разумное определение $p(r)$ как "вероятности" того, что представление $10^{e_N} \cdot f_N$ "случайного" положительного целого числа N удовлетворяет неравенству $10 f_N < r$.

Для начала попробуем найти эту вероятность, используя предельный переход, аналогично тому как мы определяли "Pr" в § 3.5. Удобный способ перефразировать это определение состоит в следующем:

$$P_0(n) = \begin{cases} 1, & \text{если } n = 10^e \cdot f, \text{ где } 10f < r, \text{ т. е. если } (\log_{10} n) \bmod 1 < \log_{10} r; \\ 0 & \text{в противном случае.} \end{cases} \tag{6}$$

Итак, последовательность $P_0(1), P_0(2), \ldots$ есть бесконечная последовательность нулей и единиц, причем единицы соответствуют случаям, вносящим вклад в значение вероятности. Мы можем попытаться "усреднить" эту последовательность, положив

$$P_1(n) = \frac{1}{n} \sum_{1 \le k \le n} P_0(k). \tag{7}$$

Естественно принять $\lim_{n\to\infty} P_1(n)$ в качестве искомой "вероятности" $p(r)$; именно так мы и сделали в § 3.5.

Но в данном случае этот предел не существует. Рассмотрим, например, подпоследовательность

$$P_1(s),\ P_1(10s),\ P_1(100s),\ \ldots,\ P_1(10^n s),\ \ldots,$$

где s—некоторое вещественное число, $1 \le s \le 10$. Если $s \le r$, то мы имеем

$$\begin{aligned} P_1(10^n s) &= \frac{1}{10^n s}(\lceil r\rceil - 1 + \lceil 10r\rceil - 10 + \cdots + \lceil 10^{n-1}r\rceil - 10^{n-1} + \lfloor 10^n s\rfloor + 1 - 10^n) = \\ &= \frac{1}{10^n s}(r(1 + 10 + \cdots + 10^{n-1}) + O(n) + \lfloor 10^n s\rfloor - 1 - 10 - \cdots - 10^n) = \\ &= \frac{1}{10^n s}\left(\frac{1}{9}(10^n r - 10^{n+1}) + \lfloor 10^n s\rfloor\right) + O(n), \end{aligned} \tag{8}$$

где в десятичной записи $r = r_0.r_1r_2\ldots$. При $n \to \infty$ функция $P_1(10^n s)$ стремится, таким образом, к предельному значению $1 + (r-10)/9s$. Вычисление, проведенное выше для случая $s \le r$, можно модифицировать таким образом, чтобы оно сохранило смысл и при $s > r$; при этом $\lfloor 10^n s\rfloor + 1$ заменится на $\lceil 10^n r\rceil$, так что для $s \ge r$ мы получим предельное значение, равное $10(r-1)/9s$. [См. J. Franel Naturforschende Gesellschaft, Vierteljahrsschrift, **62** (Zürich, 1917), 286–295.]

Итак, последовательность $P_1(n)$ содержит подпоследовательность, предел которой при возрастании s от 1 до r, а затем от r до 10 сначала возрастает от $(r-1)/9$ до $10(r-1)/9r$, а затем убывает снова до $(r-1)/9$. Отсюда видно, что последовательность $P_1(n)$ не имеет предела и что $P_1(n)$ не слишком хорошее приближение к нашему предполагаемому ответу $\log_{10} r$!

Так как $P_1(n)$ ни к чему не стремится, можно попытаться еще раз использовать ту же идею, что и в (7), чтобы "усреднением" устранить эту аномаль в поведении нашей последовательности. Вообще положим

$$P_{m+1}(n) = \frac{1}{n}\sum_{1\le k\le n} P_m(k). \tag{9}$$

Тогда $P_{m+1}(n)$ будет проявлять тенденцию к более правильному поведению, нежели $P_m(n)$. Попытаемся изучить поведение $P_{m+1}(n)$ для больших n. Опыт, приобретенный нами при рассмотрении частного случая $m = 0$, подсказывает, что стоит привлечь к делу подпоследовательность $P_{m+1}(10^n s)$. Именно на этом пути мы и докажем следующий результат.

Лемма Q. *Для произвольного целого числа $m \ge 1$ и произвольного вещественного числа $\varepsilon > 0$ найдутся такие функции $Q_m(s)$, $R_m(s)$ и такое целое число $N_m(\varepsilon)$, что при $n > N_m(\varepsilon)$ и $1 \le s \le 10$ выполняются неравенства*

$$\begin{aligned} |P_m(10^n s) - Q_m(s)| &< \varepsilon, \text{ если } s \le r, \\ |P^m(10^n s) - (Q_m(s) + R_m(s))| &< \varepsilon, \text{ если } s > r. \end{aligned} \tag{10}$$

Далее, функции $Q_m(s)$, $R_m(s)$ удовлетворяют соотношениям

$$\begin{aligned} Q_m(s) &= \frac{1}{s}\left(\frac{1}{9}\int_1^{10} Q_{m-1}(t)\,dt + \int_1^s Q_{m-1}(t)\,dt + \frac{1}{9}\int_r^{10} R_{m-1}(t)\,dt\right); \\ R_m(s) &= \frac{1}{s}\int_r^s R_{m-1}(t)\,dt; \\ Q_0(s) &= 1, \quad R_0(s) = -1. \end{aligned} \tag{11}$$

Доказательство. Рассмотрим функции $Q_m(s)$, $R_m(s)$, определенные формулами (11), и положим

$$S_m(t) = \begin{cases} Q_m(t), & t \le r, \\ Q_m(t) + R_m(t), & t > r. \end{cases} \tag{12}$$

Докажем лемму индукцией по m.

Пусть сначала $m = 1$; тогда $Q_1(s) = (1/s)(1 + (s-1) + (r-10)/9) = 1 + (r-10)/9s$ и $R_1(s) = (r-s)/s$. Из (8) находим, что

$$|P_1(10^n s) - S_1(s)| = O(n)/10^n;$$

это доказывает лемму при $m = 1$.

При $m > 1$ имеем

$$P_m(10^n s) = \frac{1}{s}\left(\sum_{0\le j<n}\frac{1}{10^{n-j}}\sum_{10^j\le k<10^{j+1}}\frac{1}{10^j}P_{m-1}(k) + \sum_{10^n\le k\le 10^n s}\frac{1}{10^n}P_{m-1}(k)\right).$$

Мы хотим оценить эту величину. Разность

$$\left|\sum_{10^j\le k\le 10^j q}\frac{1}{10^j}P_{m-1}(k) - \sum_{10^j\le k\le 10^j q}\frac{1}{10^j}S_{m-1}\left(\frac{k}{10^j}\right)\right| \tag{13}$$

меньше $(q-1)\varepsilon$, когда $1 \le q \le 10$ и $j > N_{m-1}(\varepsilon)$, а поскольку функция $S_{m-1}(t)$ непрерывна и потому интегрируема по Риману, то разность

$$\left|\sum_{10^j\le k\le 10^j q}\frac{1}{10^j}S_{m-1}\left(\frac{k}{10^j}\right) - \int_1^q S_{m-1}(t)\,dt\right| \tag{14}$$

меньше ε для всех j, больших некоторого числа N, не зависящего от q. Мы можем выбрать N большим, чем $N_{m-1}(\varepsilon)$. Следовательно, при $n > N$ разность

$$\left|P_m(10^n s) - \frac{1}{s}\left(\sum_{0\le j<n}\frac{1}{10^{n-j}}\int_1^{10} S_{m-1}(t)\,dt + \int_1^s S_{m-1}(t)\,dt\right)\right| \tag{15}$$

ограничена величиной

$$\sum_{0\le j\le N}\frac{M}{10^{n-j}} + \sum_{N<j<n}\frac{10\varepsilon}{10^{n-j}} + 10\varepsilon,$$

если M при всех j служит верхней границей для суммы **(13)**+**(14)**. Наконец, сумма $\sum_{0\le j<n}(1/10^{n-j})$, фигурирующая в **(15)**, равна $(1-1/10^n)/9$. Поэтому разность

$$\left|P_m(10^n s) - \frac{1}{s}\left(\frac{1}{9}\int_1^{10} S_{m-1}(t)\,dt + \int_1^s S_{m-1}(t)\,dt\right)\right|$$

становится меньше $(10/9)(10\varepsilon)$ при достаточно больших n. Сопоставляя это с **(10)** и **(11)**, видим, что все доказано. ∎

Суть леммы **Q**—утверждение о существовании предела

$$\lim_{n\to\infty} P_m(10^n s) = S_m(s). \tag{16}$$

Однако ввиду того что функция $S_m(s)$ не остается постоянной при изменении s, предел

$$\lim_{n\to\infty} P_m(n),$$

который мог бы быть нашей желанной "вероятностью", не существует ни для какого m. О создавшейся ситуации дает представление

Picture: Рис. 5. Вероятность того, что старшая значащая цифра равна 1.

рис. 5, изображающий значения $S_m(s)$ для малых m и $r = 2$.

Хотя функции $S_m(s)$ и не постоянны, так что у $P_m(n)$ не существует предела, мы видим из рис. 5, что уже для $m = 3$ значение $S_m(s)$ остается все время очень близким к $\log_{10} 2 = 0.30103\ldots$. Следовательно, у нас есть серьезные основания предполагать, что функция $S_m(s)$ очень близка к $\log_{10} r$ для больших m и даже что последовательность функций $\langle S_m(s)\rangle$ равномерно сходится к постоянной функции $\log_{10} r$.

Интересно доказать эту гипотезу явным вычислением $Q_m(s)$ и $R_m(s)$ для всех m, что и делается в доказательстве следующей теоремы.

Теорема F. *Для всякого $\varepsilon > 0$ найдется такое число N, что*

$$|P_m(n) - \log_{10} r| < \varepsilon \tag{17}$$

при $m, n > N$.

Доказательство. Ввиду леммы Q, этот результат будет доказан, если мы сможем показать, что существует такое число M, зависящее от ε, что для всех s из интервала $1 \le s \le 10$ и всех $m > M$ справедливы неравенства

$$|Q_m(s) - \log_{10} r| < \varepsilon \text{ и } |R_m(s)| < \varepsilon. \tag{18}$$

Значение R_m нетрудно определить из рекуррентной формулы (11). В самом деле, имеем $R_0(s) = -1$, $R_1(s) = -1 + r/s$, $R_2(s) = -1 + (r/s)(1 + \ln(s/r))$ и вообще

$$R_m(s) = -1 + \frac{r}{s}\left(1 + \frac{1}{1!}\ln\left(\frac{s}{r}\right) + \frac{1}{2!}\left(\ln\left(\frac{s}{r}\right)\right)^2 + \cdots + \frac{1}{(m-1)!}\left(\ln\left(\frac{s}{r}\right)\right)^{m-1}\right). \tag{19}$$

Для значений s из указанного интервала эта функция равномерно сходится к

$$-1 + (r/s)\exp(\ln(s/r)) = 0.$$

Рекуррентная формула (11) для Q_m принимает вид

$$Q_m(s) = \frac{1}{s}\left(c_m + 1 + \int_1^s Q_{m-1}(t)\,dt\right), \tag{20}$$

где

$$c_m = \frac{1}{9}\left(\int_1^{10} Q_{m-1}(t)\,dt + \int_r^{10} R_{m-1}(t)\,dt\right) - 1. \tag{21}$$

формула для общего члена последовательности, определяемой рекуррентной формулой (20), также находится без труда; надо выписать сначала выражения для нескольких первых членов, сообразить, какова общая формула, и доказать ее по индукции; мы получим, что

$$Q_m(s) = 1 + \frac{1}{s}\left(c_m + \frac{1}{1!}c_{m-1}\ln s + \frac{1}{2!}(\ln s)^2 + \cdots + \frac{1}{(m-1)!}(\ln s)^{m-1}\right). \tag{22}$$

Нам остается только вычислить коэффициенты c_m, которые в силу формул (19), (21) и (22) удовлетворяют соотношениям

$$\begin{gathered} c_1 = (r-10)/9, \\ c_{m+1} = \frac{1}{9}\left(c_m \ln 10 + \frac{1}{2!}c_{m-1}(\ln 10)^2 + \cdots + \frac{1}{m!}c_1(\ln 10)^m + r\left(1 + \frac{1}{1!}\ln\frac{10}{r} + \cdots + \frac{1}{m!}\left(\ln\frac{10}{r}\right)^m\right) - 10\right). \end{gathered} \tag{23}$$

Эта последовательность кажется очень сложной, однако в действительности ее можно без труда исследовать при помощи производящих функций. Положим

$$C(z) = c_1 z + c_2 z^2 + c_3 z^3 + \ldots.$$

Ввиду равенства $10^z = 1 + z\ln 10 + z^2(1/2!)(\ln 10)^2 + \ldots$, мы заключаем, что

$$c_{m+1} = \frac{1}{10}c_{m+1} + \frac{9}{10}c_{m+1} = \frac{1}{10}\left(c_{m+1} + c_m \ln 10 + \cdots + \frac{1}{m!}c_1(\ln 10)^m\right) + \frac{r}{10}\left(1 + \cdots + \frac{1}{m!}\left(\ln\frac{10}{r}\right)^m\right) - 1$$

есть коэффициент при z^{m+1} в разложении функции

$$\frac{1}{10}C(z)10^z + \frac{rz}{10}\left(\frac{10}{r}\right)^z\left(\frac{1}{1-z}\right) - \frac{1}{1-z}. \tag{24}$$

Это условие выполняется для всех значений m, так что (24) должно равняться $C(z)$, и мы получаем явную формулу

$$C(z) = \frac{-z}{1-z}\left(\frac{(10/r)^{z-1} - 1}{10^{z-1} - 1}\right). \tag{25}$$

Чтобы завершить наш анализ, нам надо изучить асимптотические свойства коэффициентов $C(z)$. Дробь в скобках в равенстве (25) стремится при $z \to 1$ к $\ln(10/r)/\ln 10 = 1 - \log_{10} r$, откуда следует, что

$$C(z) + \frac{1 - \log_{10} r}{1 - r} = R(z) \tag{26}$$

есть аналитическая функция комплексной переменной z в круге

$$|z| < \left|1 + \frac{2\pi i}{\ln 10}\right|.$$

В частности, разложение функции $R(z)$ сходится при $z = 1$, так что ее коэффициенты стремятся к нулю. Это показывает, что коэффициенты функции $C(z)$ ведут себя как коэффициенты функции $(\log_{10} r - 1)/(1 - z)$, так что

$$\lim_{m\to\infty} c_m = \log_{10} r - 1.$$

Наконец, сопоставляя этот результат с формулой (22), получаем, что $Q_m(s)$ стремится к

$$1 + \frac{\log_{10} r - 1}{s}\left(1 + \ln s + \frac{1}{2!}(\ln s)^2 + \ldots\right) = \log_{10} r$$

равномерно на отрезке $1 \le s \le 10$. ∎

Итак, мы доказали прямым вычислением наш логарифмический закон для целых чисел, причем одновременно обнаружили, что, хотя он служит очень хорошим приближением для описания усредненного поведения, в точности он никогда не достигается. Аналогичные результаты для других: распределений были опубликованы У. Фарри и Х. Гурвицем [Nature, **155** (Jan. 13, 1945), 52–53]. Доказательства леммы Q и теоремы F, которые были здесь приведены, представляют собой упрощенный и обобщенный вариант рассуждении, принадлежащих Бетти Джин Флехингер [*AMM,* **73** (1966), 1056–1061]. Другой интересный подход к распределениям, связанным с плавающей точкой, был предложен Эланом Г. Конхеймом [*Math. Comp.,* **19** (1965), 143–144].

Упражнения

1. [13] Если u и v—десятичные числа с плавающей точкой, имеющие один и тот же знак, то каково, согласно таблицам Суини, приближенное значение вероятности того, что при вычислении значения $u \oplus v$ произойдет переполнение дробной части?
2. [40] Проведите дальнейшие эксперименты со сложением и вычитанием чисел с плавающей точкой для уточнения таблиц Суини.
3. [15] Найдите, исходя из логарифмического закона, вероятность того, что две начальные цифры десятичного числа с плавающей точкой суть "23".
4. [18] В тексте отмечено, что начальные страницы интенсивно используемых таблиц логарифмов потрепаны в большей степени, чем последние страницы. А если бы мы работали вместо этого с таблицей *антилогарифмов,* т. е. таблицей, которая для данного значения $\log_{10} x$ указывает значение x, какие страницы были бы тогда самыми потрепанными?

>5. [M20] Предположим, что вещественное число U равномерно распределено в интервале $0 < U < 1$. Каково распределение наиболее значимой цифры U?

6. [22] Если бы одно слово двоичной ЭВМ содержало $n+1$ битов, то мы могли бы использовать p битов для представления дробной части двоичных чисел с плавающей точкой, один бит для знака и $n - p$ битов для показателя. Это означает, что интервал изменения представимых значений, т. е. отношение наибольшего положительного нормализованного значения к наименьшему, по существу равен $2^{2^{n-p}}$. То же машинное слово можно было бы использовать и для представления *шестнадцатеричных* чисел с плавающей точкой, выделив $p + 2$ битов для дробной части ($(p + 2)/4$ шестнадцатеричных цифр) и $n - p - 2$ битов для показателя; тогда интервал изменения значений был бы $16^{2^{n-p-2}} = 2^{2^{n-p}}$, т. е. тот же, что и раньше, причем с большим числом битов в дробной части. Может возникнуть впечатление, что мы получили что-то из ничего, однако условие нормализации в случае основания 16 слабее в том смысле, что дробная часть может содержать нули в трех наиболее значимых битах; таким образом, не все из $p+2$ битов "значащие".

Исходя из логарифмического закона, выясните, какова вероятность того, что дробная часть положительного нормализованного шестнадцатеричного числа с плавающей точкой имеет в точности 0, 1, 2 и 3 нулевых наиболее значимых бита? Основываясь на материале, изложенном в этом пункте, обсудите вопрос о достоинствах шестнадцатеричной системы в сравнении с двоичной.

7. **[BM28]** Докажите, что не существует функции распределения $F(u)$, удовлетворяющей соотношению **(5)** для каждого целого числа $b \geq 2$ и для всех вещественных значений r из интервала $1 \leq r \leq b$.
8. **[M23]** Выполняется ли соотношение **(10)** при $m = 0$ для соответствующим образом выбранного $N_0(\varepsilon)$?
9. **[BM24]** Пусть $\langle x_n \rangle$—ограниченная последовательность вещественных чисел, такая, что предел $\lim_{n\to\infty} x_{\lfloor 10^n s \rfloor} = q(s)$ существует для всех s из интер-