Учебно-исследовательский семинар «Распределение первых значащих цифр»

А. И. Щетников, А. В. Щетникова

ВВЕДЕНИЕ

Настоящая работа продолжает цикл исследований и разработок в области методики преподавания математики, связанных с проработкой содержания общего математического образования школьников и адекватных этому содержанию форм ([4], [5]). В своей работе мы исходим из того положения, что математика будет восприниматься учащимися как действительно осмысленная дисциплина, если они увидят, каким образом создаются математические понятия, и если у них будет воспитано умение математически исследовать явления реального мира.

В качестве формы такого осмысленного освоения математики могут использоваться учебно-познавательные семинары по математике для старшеклассников [6]. Участники этих семинаров, как правило, не учатся в специализированных школах и классах. Для большинства из них интерес к математике является скорее частью общего широкого интереса к миру природы и человеческой деятельности, нежели стремлением к профессиональной специализации, предполагающей владение математикой (хотя одно не исключает другое). Для работы с такими школьниками нужны задачи, которые были бы общепонятными по формулировке и богатыми по содержанию; эти задачи должны допускать простые и в то же время разнообразные подходы к их решению.

Задача, которая обсуждалась на семинарах, прошедших 22–24.03.2002 в Центре образования города Междуреченска Кемеровской области, а затем 11–13.04.2002 в школе №42 города Кемерово, является на наш взгляд одной из самых удачных в отношении этих критериев. Школьникам было предложено развернуть исследование эмпирического закона, которому подчиняется распределение первых значащих цифр в разнообразных массивах числовых данных. Эта статистическая закономерность была впервые обнаружена в 1881 г. Саймоном Ньюкомбом [10] и затем переоткрыта в 1938 г. Фрэнком Бенфордом [7], по имени которого её принято называть; согласно легенде, оба автора заметили, что в таблицах логарифмов, которыми они пользовались, особенно много следов чтения хранили первые страницы книги — те, на которых помещались логарифмы чисел, начинающихся с единицы. Объяснению и моделированию закона Бенфорда посвящён ряд публикаций в научной и научно-популярной литературе (см. обзоры [1], [3], [9], [12], [13]; обширная библиография по данной теме приведена в [8]). В последние годы закон Бенфорда стал превращаться из математического курьёза в инструмент для исследований; наиболее известно предложение Марка Негрини применять закон Бенфорда в качестве детектора лжи при аудиторских проверках [11].

–

ПЕРВЫЙ ДЕНЬ

Постановка задачи. Откроем энциклопедический словарь на произвольной странице, а затем станем перебирать подряд все числа, которые нам встретятся в словарных статьях, и будем смотреть, на какую значащую цифру они начинаются. Как вы думаете, какая первая цифра будет попадаться чаще всего? Нетрудно догадаться, что это будет цифра «1»: ведь значительная доля чисел в словаре — это чьи-то годы жизни, а большинство тех, о ком упоминает словарь, жили в одна тысяча каком-то году. Ну а если мы условимся пропускать даты? Нам будут встречаться данные о населении городов, площадях островов и стран, хозяйственные показатели и разные другие величины. Как теперь распределятся числа по их первым значащим цифрам? Будет ли их поровну, или на одни цифры будет начинаться больше чисел, чем на другие?

Подготовка массива числовых данных. В качестве источника использовались тома Большой советской энциклопедии и Детской энциклопедии, несколько общих и специализированных энциклопедических словарей, справочная часть атласа мира. За 1 час работы коллектив в 40 человек может обработать массив в 25–30 тысяч чисел. На рис. 1 показано распределение первых цифр в суммарном массиве в 53270 чисел, обработанном участниками семинаров в Междуреченске и Кемерово.

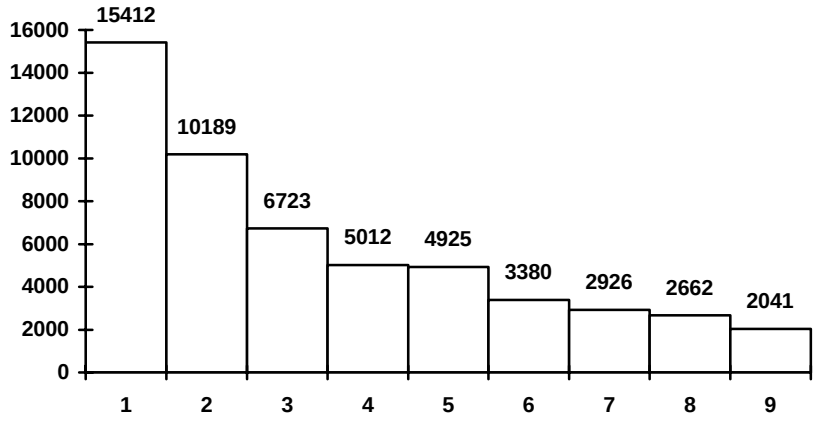


Рис. 1

Исходные вопросы. После того, как все данные были сведены в один массив, участники семинара в общей дискуссии наметили основные темы дальнейшей работы. Во-первых, нужно объяснить, почему больше всего чисел начинается на цифру «1» и меньше всего на цифру «9»; во-вторых, желательно установить математическую формулу, описывающую зависимость количества чисел от первой цифры. Затем, работая в группах по 6–8 человек, школьники искали ответы на эти вопросы.

ВТОРОЙ ДЕНЬ

Второй день начался с общего заседания, на котором обсуждались результаты групповой работы. Затем итоги общего заседания были обсуждены по группам, и каждый

участник семинара выделил то направление дальнейших исследований, которым он хотел бы заняться. День завершился разделением участников по новым группам в соответствии с выбранными направлениями. Ниже рассмотрены версии, выдвинутые группами на общем заседании.

А. Было предложено несколько вариантов объяснения обнаруженной зависимости.

(1) То, что такую большую долю первых цифр составляют единицы, можно попытаться объяснить тем, что многие числа в энциклопедии приведены в сильно округлённом виде («Этим горам 10 млн. лет»). Но это никак не объясняет, почему так много начальных двоек и троек.

(2) Счёт всегда начинается с единицы. И в каждом списке порядковых числительных (крестовые походы — с первого по восьмой, симфонии Чайковского — с первой по шестую, и т. п.) обязательно будут встречаться номер «1». Двоек будет меньше (например, в России был император Павел I, и не было Павла II), троек будет ещё меньше, и т. д. Эту гипотезу можно проверять на распределении номеров домов в случайно выбранных почтовых адресах. Впрочем, следует заметить, что большая часть данных в энциклопедии — это не порядковые, а количественные числительные.

(3) Имеет ли какое-то отношение к делу наша десятичная система счисления? Если взять наш массив данных, и пересчитать эти данные в какую-нибудь другую систему счисления, сохранится ли вид распределения?

(4) Может быть, всё дело в подсознательной избирательности человеческого мозга? Люди включали в энциклопедию такие данные, какие им больше нравятся. Возможно, что числа, начинающиеся с «1», почему-то нравятся больше. Идея экспериментальной проверки: возьмём школьный учебник математики, где нет никаких величин — результатов реального измерения, а все числа придуманы автором, и посмотрим, какое в нём будет распределение чисел по первым цифрам.

(5) Чем больше нечто, тем оно реже встречается. На Земле мало крупных животных, много мелких. Мало больших рек, много ручьёв. На одну реку длиной от 900 до 1000 км придётся, быть может, десяток рек длиной от 100 до 200 км. Описанное положение дел можно в некоторых случаях объяснить конкуренцией, возникающей вокруг ограниченного ресурса. К примеру, рассмотрим следующее модельное распределение стран по площади. Примем площадь всей суши за единицу. Пусть имеется одна страна с площадью 1/2. На остатке суши располагается одна страна с площадью 1/4, на остатке — ещё одна страна с площадью 1/8, и т. д. Какое при этом получается распределение первых цифр, можно проверить в вычислительном эксперименте. Конечно, принятая модель выглядит слишком примитивной; но её можно видоизменить так, чтобы она была ближе к реальности.

(6) Многие процессы развития сначала идут медленно, а потом всё быстрее и быстрее. К примеру, пусть в каком-нибудь посёлке живёт 10 тысяч человек. Пока он разрастётся до 20 тысяч, пройдёт довольно много времени, и всё это время первая цифра ко-

личества его жителей будет «1». Рост от 20 до 30 тысяч жителей потребует меньше времени, от 30 до 40 тысяч — ещё меньше времени, и т. д. Когда город достигнет 100 тысяч жителей, начнётся новый виток развития: снова много времени потребуется, чтобы он разросся до 200 тысяч, меньше — от 200 до 300 тысяч, ещё меньше — от 300 до 400 тысяч, и т. д. Если мы посмотрим *одновременно на разные города*, то окажется, что значительная их часть будет находиться на первом этапе очередного витка развития с первой цифрой «1», меньшая — на следующем этапе с первой цифрой «2», и т. д.

(7) Очень важный пример — таблица фундаментальных физических величин в физическом энциклопедическом словаре. Видно даже без подсчёта, что очень много чисел начинается в ней со значащей единицы. Все они приводятся с несколькими значащими цифрами. Стало быть, округление здесь ни при чём. Ничего здесь не объяснишь и избирательностью мозга. Кроме того, попавшие в таблицу величины совершенно разнородны: подряд идут скорость света, число Авогадро, масса протона, заряд электрона, магнетон Бора, постоянная Больцмана и т. п. Поэтому здесь не проходит ни объяснение с тем, что «больших величин мало, маленьких много», ни объяснение, основанное на идее экспоненциального роста.

Б. Какая математическая формула хорошо подходит для описания получившегося распределения? Школьникам известны две основные зависимости с похожим графиком: это геометрическая прогрессия и обратная пропорциональность (график — гипербола). В геометрической прогрессии должно быть $F(n + 1) : F(n) = \text{const}$, а у рассматриваемого распределения это отношение заметно убывает от первых цифр к последним. Рассмотрим теперь предположение о том, что точки лежат на гиперболе $nF(n) = \text{const}$. Для проверки построим таблицу произведений $nF(n)$:

n	1	2	3	4	5	6	7	8	9
$F(n)$	15412	10189	6723	5012	4925	3380	2926	2662	2041
$nF(n)$	15412	20378	20169	20048	24625	20280	20482	21296	18369

Среднее арифметическое $nF(n)$ равно 20118. По сравнению с этим средним значением сильно «завалены» цифры «1» и «9»; кроме того, наблюдается явный избыток пятёрок. Завал цифры «9» можно объяснить округлением девяток до десяток. Этой же причиной объясняется и избыток пятёрок: в часто встречающихся выражениях типа «сила тока увеличивается в 5–10 раз» число «5» является результатом приближённой оценки, замещающим и четвёрку, и шестёрку. Но чем объяснить завал единиц, ведь их тогда должно быть ещё больше? Похоже на то, что искомая функция ведёт себя как гипербола при больших n, но заметно отличается от неё при малых n.

В. Когда каждый участник строил график для своего массива данных (содержавшего в среднем 700 чисел), на этом графике зачастую имелись довольно сильные отклонения от монотонно и всё более полого убывающей функции (шестёрок могло быть меньше, чем семёрок, и т. п.). В массивах, построенных по группам, эти отклонения сглажива-

лись. А когда мы свели вместе данные всех групп, график сгладился ещё сильнее. Правдоподобно предположить, что если бы мы обработали в 100 раз больше чисел, график имел бы ещё более гладкий вид. И именно для предельной сглаженной функции нам и надо искать математическую формулу.

Если существует формула, описывающая сглаженную функцию $F(x)$, то по этой формуле можно вычислить значения $F(n)$ для $n > 9$. Но какой смысл можно приписать этим значениям в рамках нашей задачи? Кажется правдоподобным предположение о том, что $F(10)$ — это количество чисел рассмотренной выборки, начинающихся с «10», $F(11)$ — это количество чисел, начинающихся с «11», и т. д. Но тогда получается, что должны выполняться соотношения

$$F(10) + ... + F(19) = F(1),$$

$$F(20) + ... + F(29) = F(2),$$

. . . . . .

$$F(90) + ... + F(99) = F(9).$$

Можно выписать и другие соотношения такой же природы, например

$$F(100) + ... + F(109) = F(10).$$

Для экспериментальной проверки представляет интерес распределение чисел по вторым цифрам после первой единицы. В частности, можно проверить предсказание о том, что если $F(n)$ ведёт себя как гипербола при больших n, то значение $F(10)$ должно быть примерно в 2 раза больше $F(20)$, и т. п.

Г. Возьмём две различные единицы длины A и B. Поскольку в мире нет никаких «привилегированных» длин, шансы произвольно взятой длине попасть в интервал $[A, 2A]$ из соображений отсутствия достаточного основания должны быть равны шансам попасть в интервал $[B, 2B]$. Ещё лучше будет рассматривать объединения интервалов $\bigcup_{n=-\infty}^{+\infty} [10^n \cdot A,\ 10^n \cdot 2A]$ и $\bigcup_{n=-\infty}^{+\infty} [10^n \cdot B,\ 10^n \cdot 2B]$.

Положим теперь $B = 2A$. Получается, что шансы произвольной длине попасть в интервалы $[A, 2A]$ и $[2A, 4A]$ равны между собой. Для мерки A результаты измерений длин, попавших в первый интервал, начинаются с цифры «1»; результаты измерений длин, попавших во второй интервал, начинаются с цифр «2» и «3». Отсюда следует, что с единицы должно начинаться столько же чисел, сколько с двойки и тройки, вместе взятых. Аналогичным образом выводятся ещё три соотношения, которым должна удовлетворять функция $F(n)$:

$$F(1) = F(2) + F(3); \qquad (1)$$

$$F(2) = F(4) + F(5); \qquad (2)$$

$$F(3) = F(6) + F(7); \qquad (3)$$

$$F(4) = F(8) + F(9). \qquad (4)$$

Из соотношений (1)–(4) можно вывести разнообразные следствия. К примеру, сложив почленно первые две, либо первые три, либо все четыре формулы, мы получим

$$F(1) = F(3) + F(4) + F(5) = F(3) + F(4) + F(5) + F(6) = F(5) + F(6) + F(7) + F(8) + F(9).$$

Этим соотношениям может быть дана простая интерпретация: все числа, которые начинались с «1», при увеличении мерки в 3 раза будут начинаться с цифр «3», «4», «5»; при увеличении мерки в 4 раза — с цифр «4», «5», «6», «7», при увеличении мерки в 5 раз — с цифр «5», «6», «7», «8», «9».

Сразу же можно проверить, насколько хорошо формулы (1)–(4) описывают наше распределение:

$F(1) = 15412$, $F(2) + F(3) = 16912$ (+10%);

$F(2) = 10189$, $F(4) + F(5) = 9937$ (–2,5%);

$F(3) = 6723$, $F(6) + F(7) = 6306$ (–6%);

$F(4) = 5012$, $F(8) + F(9) = 4703$ (–6%).

Ещё одно важное следствие соотношений (1)–(4) состоит в том, что если искомая функция $F(n)$ удовлетворяет этим соотношениям, то она не является обратной пропорциональной зависимостью. Ведь в случае обратной пропорциональности $F(2) = F(1)/2$, $F(3) = F(1)/3$; но тогда $F(2) + F(3) \neq F(1)$.

ТРЕТИЙ ДЕНЬ

А. Экспериментальная группа №1 исследовала, что произойдёт с распределением первых цифр при переводе результатов в восьмеричную систему счисления. Была обработана выборка в 700 чисел. Распределение в целом сохранило тот же самый вид, оказавшись инвариантным по отношению к выбору основания системы счисления.

Б. Экспериментальная группа №2 исследовала распределение вторых цифр после единицы, стоящей на первом месте. Была обработана выборка в 3786 чисел. Полученное распределение (рис. 2) хорошо совпадало с теоретическим предсказанием, за исключением очень сильно завышенного сочетания начальных цифр «10». По-видимому, группа нарушила договорённость брать только такие данные, которые заведомо не округлялись до первой значащей цифры (пример данных, приведённых с недостаточной точностью: «В Португалии в 1982 г. производилось 10 млн. гектолитров вина»).

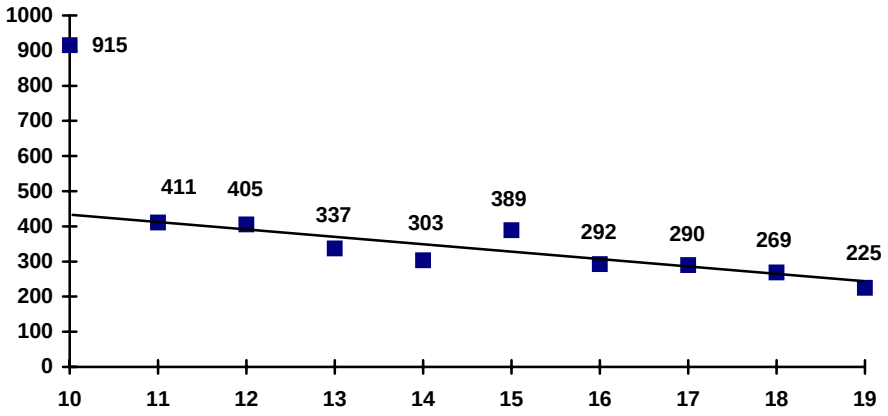


Рис. 2

Прямая $P(x) = 639 - 21x$, проведённая методом наименьших квадратов для точек с n от 11 до 19, даёт значения $P(10) = 429$, $P(20) = 219$, что неплохо согласуется с предсказанием $P(20) \approx P(10)/2$.

В. Теоретическая группа №1 пыталась установить явный вид функции $F(n)$, удовлетворяющей соотношениям (1)–(4). Рассмотрим переменную величину $G(t)$, растущую по показательному закону. Время, за которое $G(t)$ возрастает от 1 до 10, примем за единицу времени; тогда $G(t) = 10^t$. Разделим интервал [0, 1] на отрезки, внутри которых значения $G(t)$ заключены между последовательными целыми числами. Их границами служат точки lg1 = 0, lg2, lg3, ... , lg9, lg10 = 1 (рис. 3).

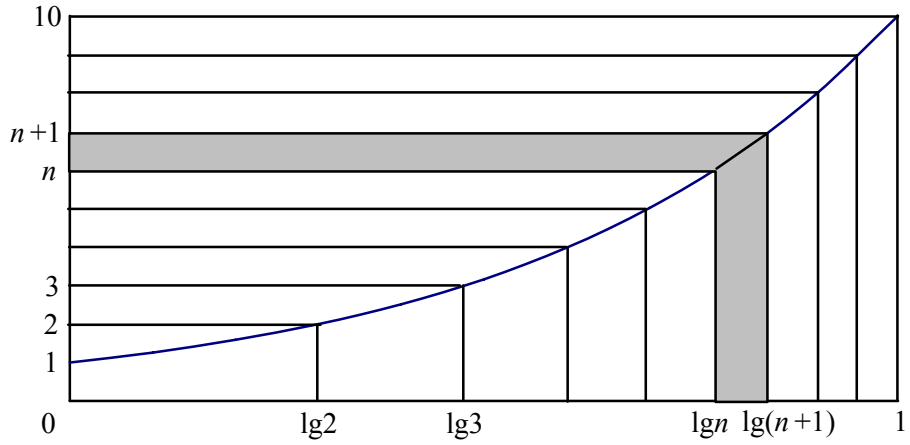


Рис. 3

Когда $G(t)$ нарастёт до 10, примем эту десятку за новую единицу измерения, а текущее время — за новое начало отсчёта; при этом процесс нарастания $G(t)$ в следующем разряде от новой единицы до новой десятки каждый раз будет описываться одной и той же формулой.

Вероятность обнаружить величину G в таком состоянии, что её первая цифра равна «n», равна длине n-ого отрезка:

–

$$F(n) = \lg(n+1) - \lg(n) = \lg\left(\frac{n+1}{n}\right) = \lg\left(1+\frac{1}{n}\right). \qquad (5)$$

Значения $F(n)$, вычисленные по формуле (5), приведены в таблице:

n	1	2	3	4	5	6	7	8	9
$F(n)$	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

Для сравнения изобразим на одном графике (рис. 4) как исходные экспериментальные данные, так и результаты, предсказываемые теорией (так называемый «логарифмический закон Бенфорда»):

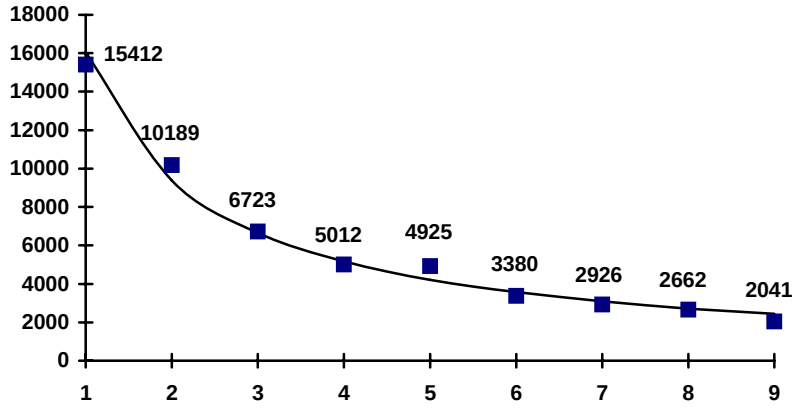


Рис. 4

Г. Ближайшее следствие рассуждений предыдущего пункта состоит в том, что формулой (5) описывается распределение первых цифр последовательных степеней любого числа a, десятичный логарифм которого иррационален, например, степеней двойки:

1, 2, 4, 8, 1, 3, 6, 1, 2, 5, 1, 2, 4, 8, ...

Геометрической прогрессии значений $G(t)$ со знаменателем a будет соответствовать арифметическая прогрессия значений t с разностью $\lg a$. Если замкнуть отрезок [0, 1] в окружность, последовательность точек, отложенных на этой окружности в одном направлении с постоянной иррациональной разностью, распределится по окружности равномерно [1]. Строгое доказательство этой теоремы, данное Г. Вейлем, довольно сложно, но легко понять её справедливость, построив достаточно большое число точек.

Д. Экспериментальная группа №3 исследовала «влияние человеческого мозга на распределение первых цифр». С этой целью были взяты несколько разных учебников математики за 6–8 классы (Междуреченск) и 2–3 классы (Кемерово), написанных разными авторами. Ясно, что числа в таком учебнике не являются результатами каких-то реальных измерений, но выдуманы авторами по их прихоти. Результаты, полученные на семинаре в Кемерово (обработан массив в 3458 чисел) и их сравнение с теоретическим

распределением (5) приведены на рис. 5. Здесь сохраняется тенденция «больше всего единиц, и меньше всего девяток», хотя и не в столь выраженном виде, как это было с числами из энциклопедии.

Как можно объяснить это явление? Предположим, что когда автор учебника хочет написать какое-нибудь двузначное число, он подсознательно выбирает с равной вероятностью, будет ли оно «небольшим», «средним» или «большим». Если «небольшими» являются числа от 10 до 25, «средними» — от 26 до 55, а «большими» — от 56 до 99, то получится нечто похожее на наши экспериментальные данные.

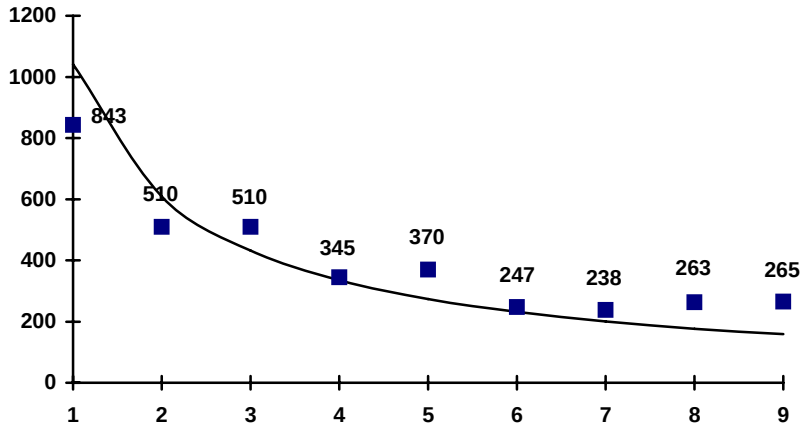


Рис. 5

Заметные отклонения от теоретической зависимости (5) можно объяснить тем, что автор учебника действует отчасти сознательно, подбирая «удобные» числа. Чтобы рассмотреть явление случайного выбора в чистом виде, был проведён опрос, в ходе которого респондентам (не знавшим о цели опроса) предлагалось назвать несколько чисел. Было опрошено 65 человек, назвавших 385 чисел. Результаты опроса и их сравнение с теоретической зависимостью (5) приведены на рис. 6. Завышенное число девяток можно объяснить «повышением внимания» к завершающим числам периода.

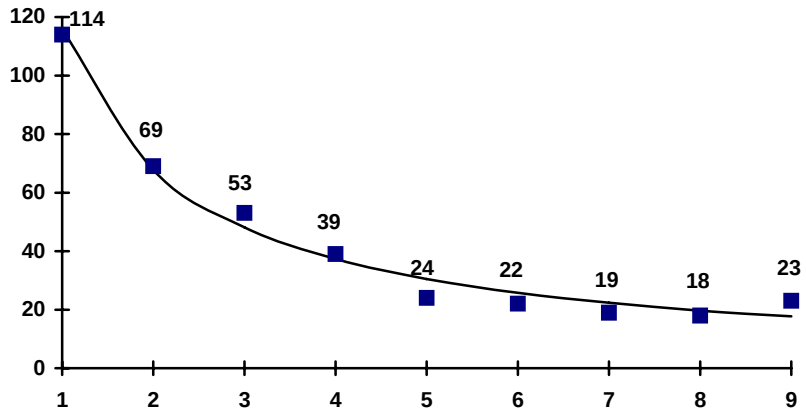


Рис. 6

Е. Теоретическая группа №2 рассматривала и обсчитывала модель речного бассейна. Пусть имеется одна река, длину которой мы примем за единицу. От этой реки отходит N_1 притоков первого уровня длиной $a < 1$. От этих притоков в совокупности отходит N_2 притоков второго уровня длиной a^2. От этих притоков в совокупности отходит N_3 притоков третьего уровня длиной a^3. И т. д.

Сначала был рассмотрен простейший вариант $N_k = 1$. Тем самым задача свелась к исследованию распределения первых цифр у степеней различных чисел a, (при этом можно брать $1 < a < 10$). Были экспериментально проверены последовательные степени чисел 2, 3, 5, 7 (при обсчёте на компьютере проверялся массив в 32000 элементов). При этом было обнаружено исключительно хорошее соответствие получившегося распределения с распределением (5). (Объяснение этого факта приведено выше в пункте **Г**.)

Затем были рассмотрены ещё два варианта этой модели: $N_k = (k + 1)$ и $N_k = (k + 1)^2$. В обоих случаях итоговое распределение также было очень близко к распределению (5).

Ж. Интересно посмотреть, как ведут себя первые цифры других числовых последовательностей; к примеру, факториалов, простых чисел или $x_n = a^{2^n}$. Для последней последовательности был проведён вычислительный эксперимент, в результате которого выяснилось, что распределение первых цифр здесь также стремится к предельному распределению (5), но этот процесс сходится гораздо медленнее. Поведение первых цифр этой последовательности обсуждалось в лекции, прочитанной перед междуреченскими школьниками, пожелавшими продолжить работу над задачей о распределении первых цифр за рамками семинара (см. приложение 1).

З. Пример массива величин, явно не удовлетворяющего закону Бенфорда, представляет собой рост взрослых людей, заключённый по большей части между 150 и 190 см. Если рост измерять в метрах, почти все результаты будут начинаться с цифры «1»; если же рост измерять в футах, большая часть результатов будет начинаться с цифры «5», следующими по частоте будут цифры «4» и «6».

Поскольку далеко не все массивы числовых данных удовлетворяют распределению первых цифр (5), следует поставить вопрос о том, для каких массивов однородных данных закон Бенфорда будет выполняться, а для каких не будет. Один из возможных ответов на этот вопрос обсуждается в приложении 2.

ЗАКЛЮЧЕНИЕ

В заключительном слове участникам семинара была рассказана притча, о которой нам недавно напомнил А. В. Боровских (см. также статью [2]).

Ученик пожаловался учителю: «Как же так, я учусь и учусь, а так многого не знаю?» Учитель ответил: «Это так, но мое незнание гораздо больше твоего». Ученик спросил, как это может быть, а учитель нарисовал круг и сказал: «Внутри этого круга находится то, что человек знает; снаружи — то, чего он не знает. А граница круга — это *наше знание о том, чего мы ещё не знаем*. Круг твоего знания невелик — поэтому мала и его

граница. А у меня знаний больше, но и граница тем самым больше — я больше тебя знаю о том, чего я не знаю».

Изучая распределение чисел по первым цифрам, мы выдвинули ряд предположений о том, как и почему происходит их убывание от единицы к девятке, и проверили некоторые из этих предположений экспериментально. Мы даже подобрали математическую формулу, неплохо описывающую это распределение, и в какой-тот мере сумели обосновать её. Мы выяснили, что именно такое распределение имеют первые цифры степеней различных чисел. Но при этом мы до сих пор не знаем причины, по которой именно это распределение реализуется для величин реального мира (или всё-таки правильнее будет говорить — для данных из энциклопедии?) Ещё было бы очень полезно выяснить, для каких массивов числовых данных закономерность (5) не выполняется, и объяснить, почему. На смену прежним вопросам пришли другие. Можно сказать, круг нашего незнания расширился.

ПРИЛОЖЕНИЕ 1

Будем исследовать распределение первых цифр числовой последовательности, заданной формулой $x_n = a^{2^n}$, или, что то же самое, рекуррентной формулой $x_{n+1} = x_n^2$ с начальным условием $x_1 = a$ (достаточно считать, что $1 \le a < 10$).

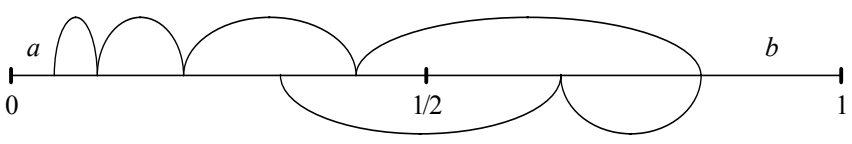
Мы уже знаем, что в задаче о распределении первых цифр удобно сделать замену $p_n = \{\lg x_n\}$ (фигурными скобками здесь обозначено взятие дробной части числа). И если точки последовательности p_n будут равномерно распределены по интервалу [0, 1), то тогда первые значащие цифры последовательности x_n буду иметь логарифмическое распределение (5).

Отображению $x_{n+1} = x_n^2$ соответствует отображение

$$p_{n+1} = \begin{cases} 2p_n, & \text{если } p_n < 1/2 \\ 2p_n - 1, & \text{если } p_n \ge 1/2 \end{cases} \qquad (6)$$

В случае $p_n \ge 1/2$ удобно сделать замену $q_n = 1 - p_n$, тогда будет $q_{n+1} = 2q_n$.

Траектория отображения (6) в общем виде напоминает игру в волейбол на площадке, ограниченной точками 0 и 1, и разделённой на два поля сеткой 1/2. Пусть начальная точка траектории отстоит на расстояние $a < 1/2$ от точки 0. Следующие точки траектории будут отстоять от точки 0 на расстояния $2a$, $4a$, ... , составляющие геометрическую прогрессию со знаменателем 2, до тех пор, пока «мяч не перелетит через сетку» и не окажется на расстоянии b от точки 1. Следующие точки траектории будут отстоять от точки 1 на расстояния $2b$, $4b$, ... , составляющие геометрическую прогрессию со знаменателем 2, до тех пор, пока «мяч вновь не перелетит через сетку» и не окажется на расстоянии c от точки 0 (рис. 7). И так далее.



–

Рис. 7

Отображение (6) имеет единственную неподвижную точку 0–1 (концы отрезка сливаются при замыкании этого отрезка в кольцо, поэтому мы говорим об одной, а не о двух неподвижных точках). Траектории, начинающиеся в рациональных точках с нечётными знаменателями, являются замкнутыми циклами. На рис. 8, *а* изображён цикл из двух точек 1/3 и 2/3; на рис. 8, *б* изображены два цикла, образованные рациональными точками со знаменателем 7. Траектории, начинающиеся в рациональных точках с чётными знаменателями, за n шагов выходят на неподвижную точку (если знаменатель имеет вид 2^n) или цикл (если знаменатель имеет вид $2^n m$, где m — нечётно).

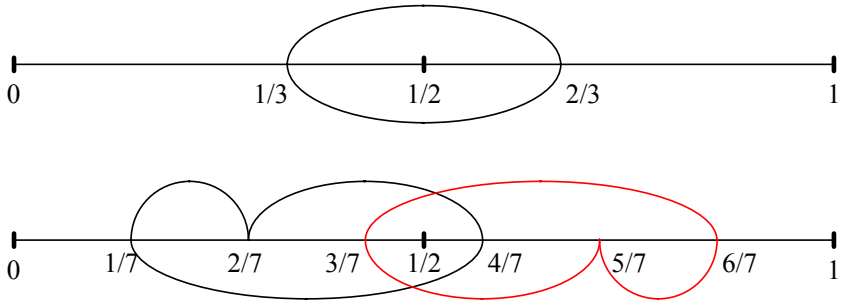


Рис. 8

Отметим тот факт, что все замкнутые циклы отображения (6), включая неподвижную точку, являются экспоненциально неустойчивыми. К примеру, рассмотрим траекторию, начальная точка которой отстоит от точки 1/3 на расстояние Δ. Далее эта траектория пройдёт по точкам

$$(1/3 + \Delta) \to (2/3 + 2\Delta) \to (1/3 + 4\Delta) \to (2/3 + 8\Delta) \to (1/3 + 16\Delta) \to \ldots ,$$

причём её удаление от точек цикла с каждым шагом растёт в геометрической прогрессии со знаменателем 2 (рис. 9).

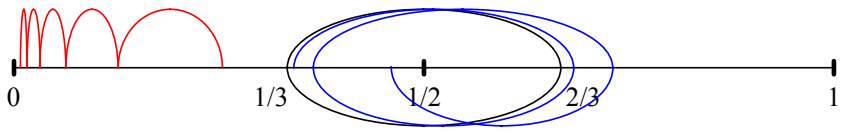


Рис. 9

Поведение траектории, начинающейся в произвольной иррациональной точке, можно качественно описать следующим образом. Траектория постоянно попадает в близкую окрестность какого-нибудь цикла с не слишком большим знаменателем, — но тут же экспоненциально «раскручивается» с этого цикла, попадает в окрестность другого цикла и вновь «раскручивается» с него, и так её «бросает» от одного цикла к другому. В численном счёте хорошо просматриваются ситуации, когда траектория попадает в близкую окрестность неподвижной точки 0–1; можно заметить также «прокручивания» траектории вблизи стационарного цикла, образованного точками 1/3 и 2/3.

Зададимся теперь вопросом, приводит ли отображение (6) при иррациональной начальной точке к равномерному распределению точек траектории по интервалу [0, 1). Численный счёт показывает, что выход на стационар действительно происходит, но этот процесс идёт относительно медленно по сравнению с аналогичным процессом для степеней двойки.

Дадим не совсем строгое доказательство того факта, что если отображение (6) при $n \to \infty$ даёт стационарное распределение точек траектории по интервалу [0, 1), то это распределение является равномерным. Рассмотрим совокупность всех точек, в которых некоторая траектория, стартовавшая с иррациональной точки, побывала за очень большое число шагов. Пусть все эти точки распределены по интервалу [0, 1) с плотностью $\rho(x)$. На следующем шаге каждая из этих точек перейдёт в другую точку интервала [0, 1) согласно (6), и при этом совокупность образов даст новое распределение плотности $\rho_+(x)$. Но поскольку мы предполагаем, что распределение $\rho(x)$ является стационарным, тем самым будет $\rho_+(x) = \rho(x)$.

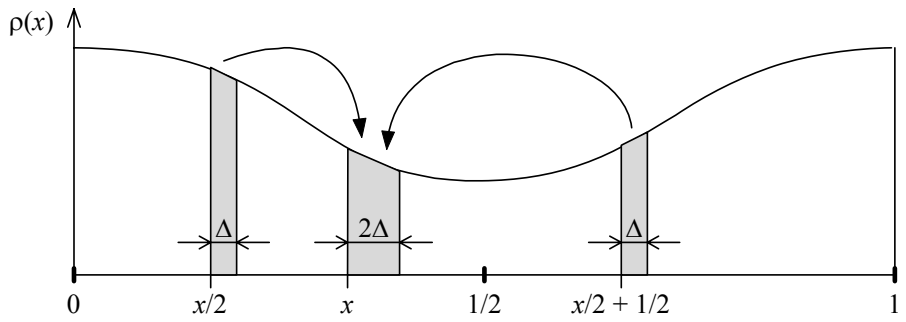


Рис. 10

Заметим теперь, что прообразами точек из интервала $[x, x + 2\Delta]$ являются все точки из интервалов $[x/2, x/2 + \Delta]$ и $[x/2 + 1/2, x/2 + 1/2 + \Delta]$ (рис. 10). Поэтому площадь криволинейной трапеции над интервалом $[x, x + 2\Delta]$ равна сумме площадей криволинейных трапеций над интервалами $[x/2, x/2 + \Delta]$ и $[x/2 + 1/2, x/2 + 1/2 + \Delta]$. Тем самым $\rho(x)$ при малых Δ должно удовлетворять соотношению

$$2\rho(x)\Delta = \rho(x/2)\Delta + \rho(x/2 + 1/2)\Delta,$$

что даёт функциональное уравнение

$$2\rho(x) = \rho(x/2) + \rho(x/2 + 1/2).$$

Непрерывные решения этого уравнения имеют вид $\rho(x) = \text{const}$. Тем самым первые цифры чисел последовательности $x_n = a^{2^n}$ будут иметь распределение (5).

ПРИЛОЖЕНИЕ 2

При объяснении эмпирического закона Бенфорда было выдвинуто предположение о том, что на каждый десятичный разряд шкалы величин приходится одно и то же число однородных объектов. Согласно этому предположению, число поселений (или стран) с численностью населения, заключённой в границах каждого десятичного разряда, должно быть одним и тем же. Очевидно, что это предположение слишком далеко от реальности. Поэтому нужно видоизменить предложенное объяснение закона Бенфорда таким образом, чтобы оно лучше соответствовало реальному положению вещей.

С этой целью рассмотрим распределение численности населения стран мира по первым цифрам внутри каждого десятичного разряда. Из таблицы видно, что избыток пятёрок в итоговом распределении по отношению к теоретически предсказанному результату обусловлен странами с численностью в пределах от 5 до 6 млн. человек.

	1	2	3	4	5	6	7	8	9	Сумма
10^4	2	4	1	3	1	2	2	1		16
10^5	8	7	2	4	2	2	1	2	1	29
10^6	11	11	12	9	13	3	4	6	4	73
10^7	29	15	5	4	5	4	4	1	1	68
10^8	6	2							1	9
10^9	1									1
Сумма	57	39	23	24	26	11	18	18	7	196
%	29,2	19,9	10,2	10,2	10,8	5,6	5,6	5,1	3,6	
% (теор.)	30,1	17,6	12,5	9,7	7,9	6,7	5,8	5,1	4,6	

На рис. 11 каждый десятичный разряд разделён на три равные части, заключённые между границами 1, $10^{1/3} = 2{,}154$, $10^{2/3} = 4{,}642$, 10. Ширина колокола функции распределения, внутрь которой попадает 50% стран, составляет один порядок шкалы.

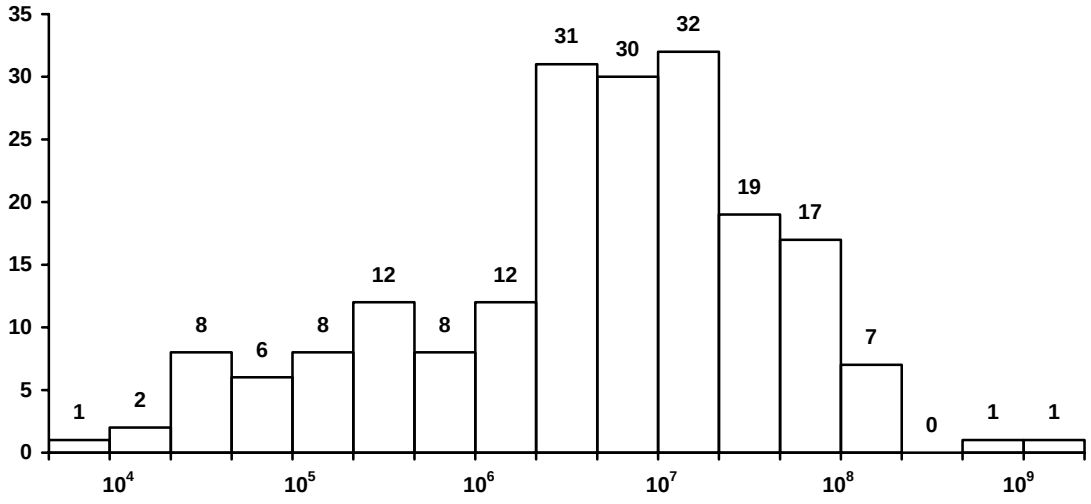


Рис. 11

Рассмотренный пример позволяет нам сформулировать основное условие, которому должны удовлетворять однородные величины реального мира, чтобы распределение их первых цифр было близко к логарифмическому закону Бенфорда. Это условие состоит в том, что ширина колокола функции их распределения по логарифмической шкале должна быть не менее одного порядка.

Чтобы получить численные оценки отклонения распределения первых цифр от закона Бенфорда (5), рассмотрим массив величин, логарифмы которых имеют нормальное распределение

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - x_0)^2}{2\sigma^2} \right).$$

Разрежем подграфик функции $\varphi(x)$ на отдельные сегменты, заключённые между соседними целыми числами, сдвинем каждый сегмент вдоль оси абсцисс так, чтобы он наложился на отрезок [0, 1], а затем найдём сумму всех подграфиков:

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{m=-\infty}^{+\infty} \exp\left(-\frac{(x - x_0 - m)^2}{2\sigma^2} \right).$$

Определённая таким образом функция $\rho(x)$ является периодической с периодом 1. Разложим её в ряд Фурье по косинусам на отрезке $[x_0 - 1/2, x_0 + 1/2]$:

$$\rho(x) = 1 + 2\sum_{n=1}^{\infty} \exp(-2(\pi n\sigma)^2) \cos(2\pi n(x - x_0)).$$

При $\sigma = 0{,}2$ амплитуда второй гармоники примерно в 10 раз меньше амплитуды первой гармоники, поэтому при $\sigma > 0{,}2$ для численных оценок можно пользоваться приближённой формулой

$$\rho(x) = 1 + 2\exp(-2(\pi\sigma)^2) \cos(2\pi(x - x_0)).$$

Функция $\rho(x)$ на отрезке [0, 1] достигает своего максимума в точке $\{x_0\}$, минимума в точке $\{x_0 + 1/2\}$. Поэтому распределение первых цифр рассматриваемого массива будет завышено по сравнению с распределением (5) в $1 + 2\exp(-2(\pi\sigma)^2)$ раз в цифре, ближайшей к $10^{\{x_0\}}$, и занижено в $1 - 2\exp(-2(\pi\sigma)^2)$ раз в цифре, ближайшей к «противоположной» цифре $10^{\{x_0+1/2\}}$.

–

ЛИТЕРАТУРА

1. Арнольд В. И. *«Жёсткие» и «мягкие» модели в математике*. М., МЦНМО, 2000.
2. Боровских А. В. Круг незнания. *Поиск*, № 10–11, 15 марта, 2002, с. 8.
3. Секей Г. *Парадоксы в теории вероятностей и математической статистике*. М., Мир, 1990.
4. Щетников А. И. Материалы к проектированию курса геометрии для средней школы. *Математическое образование*, 2000, №3(14), с. 35–42.
5. Щетников А. И., Щетникова А. В. Преподавание математики в историческом контексте. *Математическое образование*, 2001, №3(18), с. 60–68.
6. Щетников А. И., Щетникова А. В. Учебный семинар «Как решать незнакомую задачу». *Труды конференции, посвящённой 90-летию со дня рождения А. А. Ляпунова.* Новосибирск, ОИИ СО РАН, 2001, с. 773–780.
   http://www.ict.nsc.ru/ws/Lyap2001/2309/
7. Benford F. The law of anomalous numbers. *Proc. Amer. Phil. Soc.,* **78**, 1938, p. 551–572.
8. Domínguez M. P., Burguillo J. D. A. El primer dígito significativo. *Epsilon. Revista del S.A.E.M. Thales*, №45, **15(3)**, 1999, p. 339–351.
   http://www.cs.us.es/~perer/publicac/epsilon45.pdf
9. Hill T. P. The first digit phenomenon. *American Scientist,* **86**, 1998, p. 358–363.
10. Newcomb S. Note on the frequency of the use of digits in natural numbers. *Amer. J. Math.* **4**, 1881, p. 39–40.
11. Nigrini M. J. *Digital analysis using Benford's Law: tests statistics for auditors*. Global Audit Publications, 2000.
12. Raimi R. A. The peculiar distribution of first digits. *Scientific American,* **221**, Dec. 1969, p. 109–119.
13. Scott P. D., Fasli M. *Benford's Law: an empirical investigation and a novel explanation*. CSM Technical Report ¹349, University of Essex, 2001.
    http://cswww.essex.ac.uk/technical-reports/2001/CSM-349.pdf