**Закон Бенфорда или закон первой цифры**

Недавно я прочитал замечательную книгу Леонарда Млодинова [(Не)совершенная случайность. Как случай управляет нашей жизнью](http://baguzin.ru/wp/?p=4219). О-о-чень рекомендую! Некоторые фрагменты мне особо понравились, и вот сегодня об одном из них – законе Бенфорда.[[1]](#footnote-1)

Закон Бенфорда или закон первой цифры гласит, что в таблицах чисел, основанных на данных источников из реальной жизни, цифра 1 на первом месте встречается гораздо чаще, чем все остальные (рис. 1). Более того, чем больше цифра, тем меньше вероятности, что она будет стоять в числе на первом месте.



Рис. 1. Вероятность встретить первую цифру в данных, основанных на источниках из реальной жизни

Например, если подсчитать, с какой частотой встречаются первые цифры в числах, являющихся степенью двойки, то закономерность будет почти такой же (рис. 2). Аналогично ведут себя и [числа Фибоначчи](http://ru.wikipedia.org/wiki/%D0%A7%D0%B8%D1%81%D0%BB%D0%B0_%D0%A4%D0%B8%D0%B1%D0%BE%D0%BD%D0%B0%D1%87%D1%87%D0%B8) и чуть менее «красиво» [факториалы](http://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D0%BA%D1%82%D0%BE%D1%80%D0%B8%D0%B0%D0%BB) (см. лист «Рис. 2» Excel-файла).

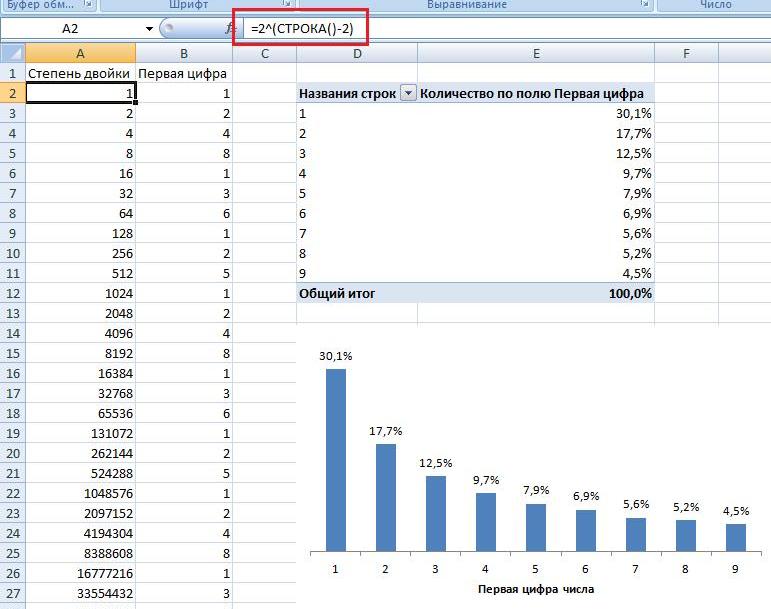


Рис. 2. Первая цифра в числах, являющихся степенью двойки, для диапазона от 20 до 21000

Закон Бенфорда был открыт вовсе не Бенфордом, а американским астрономом Шимоном Ньюкомбом. Примерно в 1881 г. Ньюкомб заметил, что страницы тетради с логарифмическими таблицами, на которых числа начинались с 1, гораздо сильнее захватаны и истрепаны, чем страницы, на которых числа начинались с 2 и так далее до 9 – те выглядели чистыми, как будто их вообще не открывали. Ньюкомб предположил: те страницы, которые больше всего истрепались, чаще всего и открывали, и на основании своих наблюдений заключил: те ученые, которые до него брали тетрадь, работали с данными, отражавшими подобное распределение цифр. Закон же был назван по фамилии Франка Бенфорда, который в 1938 г. заметил то же самое, что и Ньюкомб, когда просматривал логарифмические таблицы в научно-исследовательской лаборатории «Дженерал Электрик» в г. Скенектади, штат Нью-Йорк. Но ни Ньюкомб, ни Бенфорд не доказали справедливость закона. Это произошло только в 1995 г., и автор доказательства – Тед Хилл, математик из Технологического института Джорджии.

Закону Бенфорда подчиняются числа из многих областей, к примеру, из области финансов. В действительности, закон как нельзя лучше подходит для обработки большого массива финансовых показателей на предмет мошенничества.

В одном таком случае был замешан молодой предприниматель Кевин Лоуренс – он умудрился собрать 91 млн. долларов на создание сети клубов здоровья, оборудованных по последнему слову техники. Набив карманы наличными, Лоуренс развил бурную деятельность, нанял тучу исполнительных директоров и спустил деньги инвесторов так же быстро, как и собрал. И все бы ничего, за исключением одного: Лоуренс со своей когортой большую часть денег тратили не на развитие дела, а на личные нужды. А так как приобретение нескольких домов, двадцати личных яхт, сорока семи автомобилей (в числе которых пять «хаммеров», четыре «феррари», три спортивных «доджа», два шикарных «форда» и «ламборгини дьябло»), двух часов «Ролскс», браслета с бриллиантами в 21 карат, самурайского меча за 200 тыс. долларов и машины для коммерческого производства сладкой ваты едва ли можно было списать как деловые расходы, Лоуренс с дружками попытались увести деньги путем перечисления их по сложной банковской схеме со счета на счет как средства то одной подставной компании, то другой – все с целью создания видимости активно расширяющегося бизнеса. На их несчастье, заподозривший неладное бухгалтер-криминалист Даррелл Доррелл составил список из более чем 70 тыс. номеров (счета и переводы) и, опираясь на закон Бенфорда, сравнил, как распределяются цифры. А распределялись они вразрез с законом. Это, конечно же, было только началом расследования, однако дальше история развивалась по известному сценарию, а развязка наступила за день до Дня благодарения 2003 г., когда Кевин Лоуренс, окруженный своими адвокатами и облаченный в светло-голубую тюремную робу, был приговорен к двадцати годам заключения без права досрочного освобождения. Налоговое управление США также изучило закон Бенфорда как способ обнаружения случаев налогового мошенничества. Один исследователь даже применил закон к данным налоговых поступлений от Билла Клинтона за тринадцать лет. Цифры распределились в соответствии с законом.

Закон Бенфорда применим к множествам чисел, которые могут расти экспоненциально (другими словами, темп роста величины пропорционален её текущему значению). Например, счета за электричество, остатки товаров на складах, цены на акции, численность населения, смертность, длины рек, площади стран, высоты самых высоких сооружений в мире.

Закон обычно не действует для распределений с заданными минимальными или максимальными значениями (список компаний с доходом от 50 000 до 100 000 долларов). Также не подходит нормальное распределение и распределения, охватывающие только один или два порядка величин (IQ взрослых). Закон Бенфорда не применим к множеству букв (но, например, к словам применим [закон Ципфа](http://baguzin.ru/wp/?p=1716)). Объём данных должен быть достаточен для применения статистических методов.

Форма Закона Бенфорда может быть объяснена, если предположить, что равномерно распределены логарифмы чисел; например, вероятность нахождения числа между 100 и 1000 (логарифм между 2 и 3) является такой же, как и между 10 000 и 100 000 (логарифм между 4 и 5). Для множества чисел, особенно имеющих экспоненциальный рост, таких как доходы или цены на бирже, это разумное предположение.

Закон Бенфорда выполняется для всех процессов, имеющих фрактальную природу (см., например, [Бенуа Мандельброт. (Не)послушные рынки: фрактальная революция в финансах](http://baguzin.ru/wp/?p=1604)).

Для того чтобы установить явный вид функции F(n), удовлетворяющей закону Бенфорда, рассмотрим переменную величину G(t), растушую по показательному (экспоненциальному) закону.[[2]](#footnote-2) Время, за которое G(t) возрастает от 1 до 10, примем за единицу времени; тогда G(t) = 10t. 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).

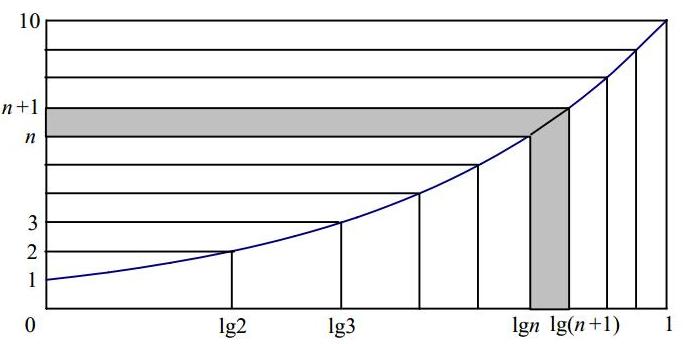
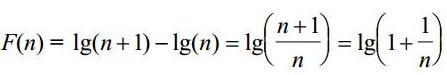


Рис. 3. Объяснение закона Бенфорда

Когда G(t) нарастёт до 10, примем эту десятку за новую единицу измерения, а текущее время – за новое начало отсчета; 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:



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

|  |  |
| --- | --- |
| Первая цифра | Значение F(n) или вероятность встретить цифру первой |
| 1 | 30,103% |
| 2 | 17,609% |
| 3 | 12,494% |
| 4 | 9,691% |
| 5 | 7,918% |
| 6 | 6,695% |
| 7 | 5,799% |
| 8 | 5,115% |
| 9 | 4,576% |

Похоже ведет себя и население стран мира (рис. 4). Вот что пишет на эту тему математик академик Владимир Игоревич Арнольд. Согласно [теории Мальтуса](http://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D0%BB%D1%8C%D1%82%D1%83%D1%81,_%D0%A2%D0%BE%D0%BC%D0%B0%D1%81), население каждой страны растет в геометрической прогрессии. А первые цифры населения фиксированной страны в последовательные годы распределены как первые цифры степеней двойки (см. рис. 2). Согласно эргодическому принципу,[[3]](#footnote-3) временное среднее можно заменить пространственным: распределение по странам в один и тот же год должно совпадать с распределением в одной стране в разные годы.

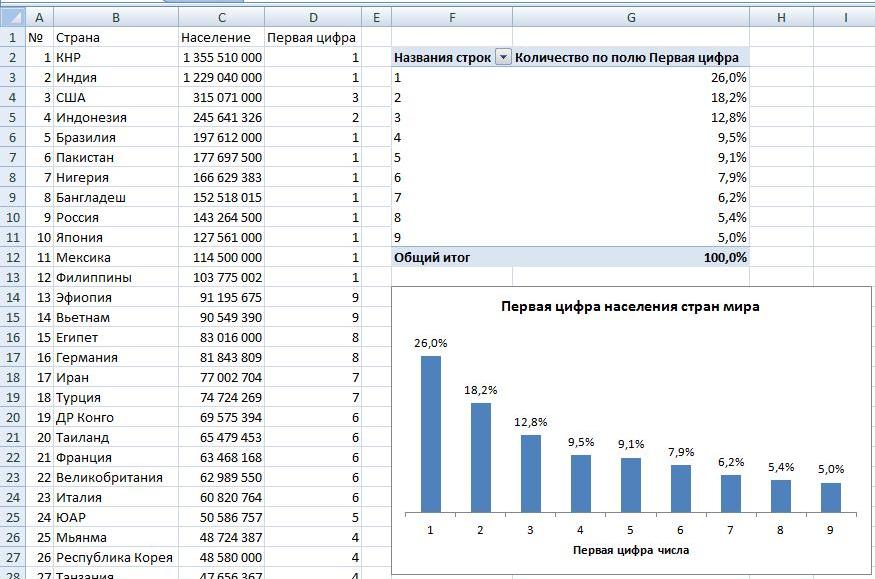


Рис. 4. Первая цифра населения стран мира

Снова предоставим слово академику Арнольду. В середине 80-х М.Б.Севрюк обнаружил, что не только населения, но и площади стран мира подчиняются такому же закону распределения первых цифр, как степени двойки (рис. 5). К площадям теория Мальтуса, по-видимому, неприменима, так что возник вопрос – как объяснить это поведение площадей. Попытаюсь дать ответ на этот вопрос.

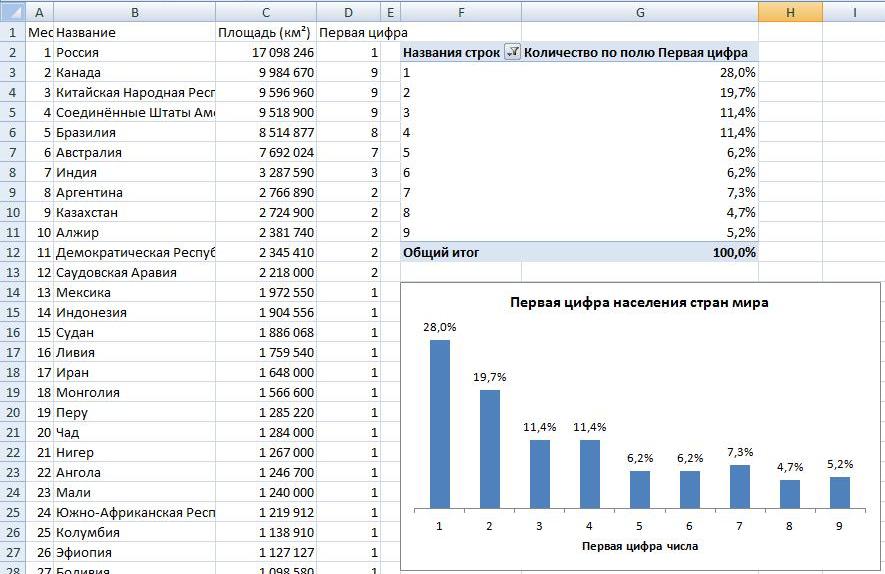


Рис. 5. Первая цифра площади стран мира

Предыдущие примеры подсказывают, что следует искать причину странного распределения первых цифр площадей стран мира, либо в их росте, либо в убывании (в геометрической прогрессии). История мира показывает, что площади стран (особенно империй) иногда растут, а иногда убывают за счет то присоединения одних стран к другим, то распада. Так что и в этом процессе просматривается геометрическая прогрессия.

В качестве примера я хотел найти цифры, как изменялась территория (площадь) России, но мне этого не удалось. Взамен предлагаю вам полюбоваться [интерактивной картой](http://ria.ru/infografika/20100902/271470994.html) изменения территории России на протяжении веков, выполненной в РИА Новости.

Следующий пример[[4]](#footnote-4) – обработка численных данных из Большой советской энциклопедии, детской энциклопедии, нескольких общих и специализированных энциклопедических словарей. За 1 час 40 школьников обработали массив в 53 270 чисел. Результаты на рис. 6.

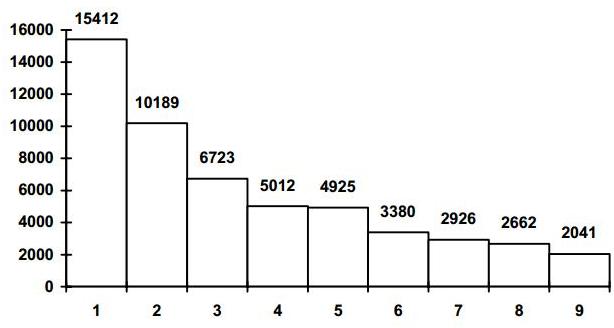


Рис. 6. Первые цифры чисел из энциклопедий

И напоследок статистика первых цифр суммы счета торговой компании (более 71 000 записей).

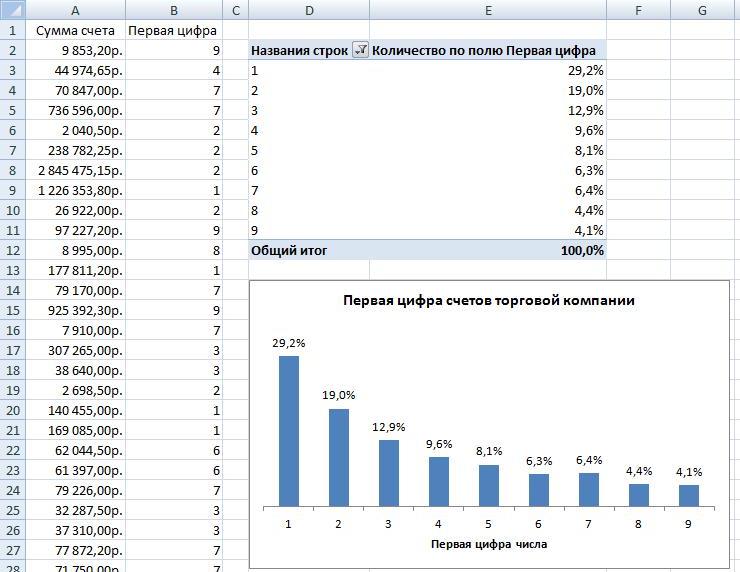


Рис. 7. Первые цифры суммы счета торговой компании

1. Заметка основана на фрагменте [книги](http://baguzin.ru/wp/?p=4219) Младинова, материалах [Википедии](http://ru.wikipedia.org/wiki/%D0%97%D0%B0%D0%BA%D0%BE%D0%BD_%D0%91%D0%B5%D0%BD%D1%84%D0%BE%D1%80%D0%B4%D0%B0), [статье](http://kvant.mccme.ru/pdf/1998/01/kv0198arnold.pdf) академика В.Арнольда, материалах А.И.Щетников, А.В.Щетникова Учебно-исследовательский семинар [Распределение первых значащих цифр](http://www.nsu.ru/classics/Benford.pdf) и собственных изысканиях [↑](#footnote-ref-1)
2. По материалам упоминавшейся ранее работы Щетниковых [↑](#footnote-ref-2)
3. [Эргодичность](http://ru.wikipedia.org/wiki/%D0%AD%D1%80%D0%B3%D0%BE%D0%B4%D0%B8%D1%87%D0%BD%D0%BE%D1%81%D1%82%D1%8C) – свойство некоторых динамических систем; для эргодических систем математическое ожидание по временным рядам должно совпадать с математическим ожиданием по пространственным рядам. [↑](#footnote-ref-3)
4. Из ранее цитированной работы Щетниковых [↑](#footnote-ref-4)