УДК 51

Доманов А.К.

школьник

Муниципальное бюджетное образовательное учреждение

средняя общеобразовательная школа №4

имени Героя Советского Союза Жукова Георгия Константиновича

муниципального образования Тимашевский район

ИСПОЛЬЗОВАНИЕ ИЗМЕНЕНИЯ СИСТЕМЫ ОТСЧЕТА ДЛЯ УЛУЧШЕНИЯ РЕЗУЛЬТАТОВ АНАЛИЗА С ПОМОЩЬЮ ЗАКОНА БЕНФОРДА

Аннотация: в данной статье предложен метод повышения эффективности анализа данных с помощью закона Бенфорда за счет предварительного перевода чисел в систему счисления с большим основанием.

В работе представлено теоретическое обоснование метода и результаты экспериментальной проверки. Показано, что такой подход увеличивает чувствительность анализа и улучшает выявление аномалий по сравнению со стандартным применением закона Бенфорда.

Ключевые слова: закон Бенфорда, анализ данных, системы счисления, статистика, закономерности.

Domanov A.K.

shkolnik

Municipal budgetary educational institution secondary general education school No. 4 named after Hero of the Soviet Union Georgy Konstantinovich Zhukov of the Timashevsky district municipality

USING A CHANGE IN THE FRAME OF REFERENCE TO IMPROVE ANALYSIS RESULTS USING BENFORD'S LAW

Abstract: This article proposes a method for improving the efficiency of data analysis using Benford's law by first converting numbers into a number system with a large base. The paper presents the theoretical justification of the method and the results of experimental verification. It is shown that this approach increases the sensitivity of the analysis and improves the detection of anomalies compared to the standard application of Benford's law.
Keywords: Benford's law, data analysis, number systems, statistics, patterns.

Закон Бенфорда, также известный как закон первой цифры, представляет собой удивительное явление, наблюдаемое в различных наборах данных, где первая цифра чисел не распределена равномерно. Согласно этому закону, в естественных наборах чисел цифра 1 появляется значительно чаще, чем цифры 2, 3 и так далее, вплоть до 9.

Саймон Ньюкомб первым заметил, что «то, что десять цифр не встречаются с одинаковой частотой, должно быть очевидно любому, кто много пользуется логарифмическими таблицами и замечает, насколько быстрее изнашиваются первые страницы, чем последние».

После него уже Фрэнк Бенфорд обратил внимание на то, что «частота первых цифр близко следует логарифмическому соотношению $F = \log\left(\frac{a+1}{a}\right)$, где F – частота цифры a на первом месте используемых чисел».

Он повсеместно используется для обнаружения мошеннических схем в финансовой сфере, проверке выборов на фальсификацию и во многих других областях, где необходимо обнаруживать аномалии в больших объемах данных. Ведь различные манипуляции с данными меняют распределение первых цифр так, что они перестают подходить под закон Бенфорда.

В этой статье рассматривается возможность перевода значений в статистике в другую систему счисления для улучшения результатов анализа.

Распределение по закону Бенфорда в современном виде выглядит так:

$$P(n)=\log_b(1+\frac{1}{n}¿)¿$$

Где: n – первая цифра какого-либо значения в распределении; b - система счисления, в которой представлено распределение; P(n) – вероятность цифры n быть первой значащей.

То есть распределение меняется при переходе из одной системы счисления в другую, что и используется в этом методе.

Для исследования как набор анализируемых данных была выбрана площадь водосборного бассейна рек мира. Всего было использовано 171 значение. Анализ проводился в системах счисления: троичной, восьмеричной, десятичной, семнадцатеричной, шестидесятеричной и от тридцатидвоичной до тридцатидевятеричной.

На графике 2 обозначены:

- **Черным** – троичная система счисления.
- **Красным** – семеричная система счисления.
- **Синим** – десятеричная система счисления.
- **Оранжевым** – семнадцатеричная система счисления.
- **Зеленым** – тридцатишестеричная система счисления.
- **Фиолетовым** – шестидесятеричная система счисления.

Пунктиром обозначена функция закона Бенфорда, какой она должна быть в идеале, а сплошной линией – функция, получившаяся в результате анализа.

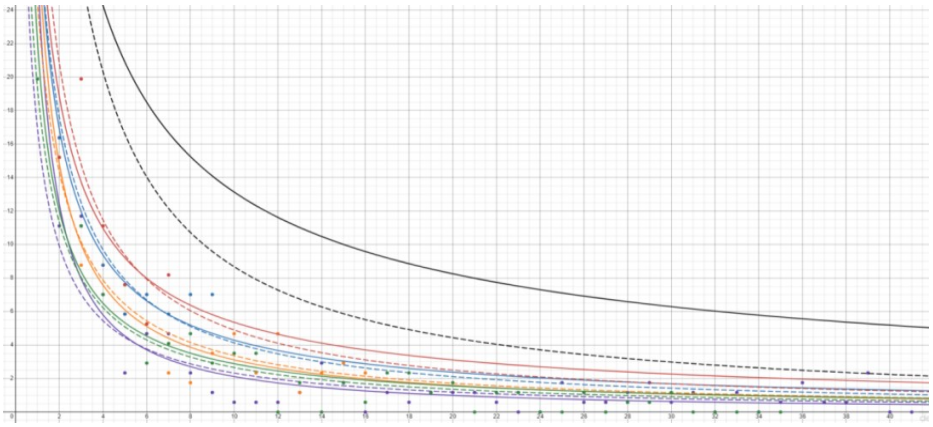

График 2. Результат вычислений.

При изучении получившейся функции можно понять, что качество анализа возрастает при увеличении системы счисления вплоть до тридцатишестеричной.

На наиболее маленьких системах счисления качество минимально, и аномалии в такой системе будут практически незаметны.

Это объясняется тем, что увеличение системы счисления увеличивает количество возможных первых цифр, «размазывая» распределение, тем самым делая аномалии более заметными.

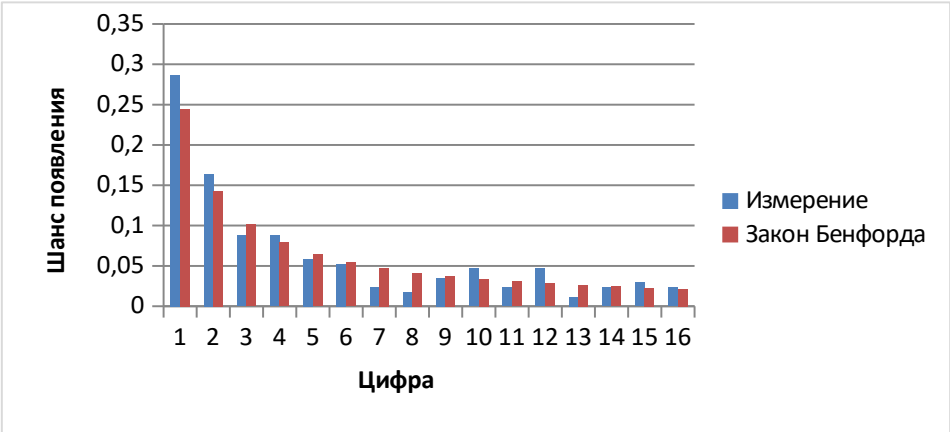

Таблица 1. Семнадцатеричная система счисления

Проблемы начинаются при значительном увеличении системы счисления, к примеру до шестидесятеричной (в тридцатишестиричной тоже иногда проглядывают искажения). При слишком большом увеличении функции становятся менее точными. Это можно объяснить тем, что основание системы счисления слишком приближается к

количеству данных, хотя аномалии в такой системе все еще будут более заметными. Когда основание системы приближается к количеству анализируемых данных, количество анализируемых цифр увеличивается, и распределяемых на них значений не хватает, что делает такой анализ ненадежным. Вдобавок если приблизится к самим значениям, нарушится условие работы закона Бенфорда об охватывании нескольких порядков величин, и он может просто перестать работать.

Таблица 2. Шестидесятеричная система счисления

В шестидесятеричной системе счисления точность становится очень маленькой (таблица 2), но зато аномалии появляются очень явно. Даже если взять большое количество данных, любое появление больших чисел означает аномалию, ведь шанс появления к примеру цифры 59 невероятно мал: 0,046%. И тем не менее точность анализа слишком низка.

То есть для того, чтобы эффективно анализировать данные, основание системы счисления b в среднем должно быть минимум в 5 раз меньше количества анализируемых данных N.

$$b \leq \frac{N}{5}$$

Также можно уменьшить систему счисления, если в десятеричной системе нарушается условие охватывания нескольких порядков величин или имеется слишком мало данных.

Вывод: у этого способа присутствуют как и плюсы, так и минусы. При использовании этого метода необходимо подобрать оптимальную систему счисления, что может занять много времени, но тем не менее он значительно увеличивает обнаружимость аномалий.

Плюсы:

- Аномалии при увеличении системы счисления становятся гораздо заметнее, так как данные размазываются и шансы встречи больших первых цифр значительно уменьшается, что также помогает при анализе.
- При подборе подходящей системы счисления точность анализа может возрасти в отличие от десятеричной системы.
- Если факт присутствия аномалии был специально скрыт в десятеричной системе, аномалия может быть обнаружена в других системах счисления.

Минусы:

- Такой анализ может быть не всегда удобным, так как происходит не в привычной для нас десятичной системе счисления. Может занять слишком много времени.
- Точность понижается при приближении к количеству данных или минимальным значениям анализируемой статистики, при слишком большом увеличении закон может вообще перестать работать.

Этот метод можно использовать либо в статистике с небольшим количеством данных для того, чтобы аномалии становились более заметными, либо в статистике с большим количеством данных для того,

чтобы сделать анализ более точным. Точность анализа повышается не всегда, в некоторых случаях может уменьшаться. Тем не менее, основное преимущество этого метода заключается в том, что все аномалии в данных будут намного более заметными в больших системах счисления.

Желательно применять закон Бенфорда при анализе, совмещая разные методы. Например можно совместить метод увеличения системы отсчета вместе с методом анализа второй цифры, то есть анализировать и первую и вторую цифру в разных системах отсчета.

Использованные источники:

1. Саймон Ньюкомб. “Note on the Frequency of Use of the Different Digits in Natural Numbers.” American Journal of Mathematics, vol. 4, no. 1, 1881, pp. 39–40. JSTOR, https://doi.org/10.2307/2369148. Accessed 9 Apr. 2025.
2. Фрэнк Бенфорд. “The Law of Anomalous Numbers.” *Proceedings of the American Philosophical Society*, vol. 78, no. 4, 1938, pp. 551–72. *JSTOR*, http://www.jstor.org/stable/984802. Accessed 9 Apr. 2025.