УДК 519.213:51-77
Научная специальность: 1.1.4; 5.2.4
https://elibrary.ru/ltxgrw

Использование распределения Бенфорда для снижения риска необнаружения искажений финансовой отчетности

ISSN 1812-5220

Криволапов С.Я.*,
Комиссарова А.В.,
Хамула Д.А.,
Финансовый университет при Правительстве Российской Федерации, 125167, Россия, г. Москва, Ленинградский пр-т, д. 49

Аннотация

Рассматривается совокупность числовых массивов, каждый из которых содержит данные о финансовой отчетности некоторых компаний. Для каждого массива определяются частоты появления каждой из возможных цифр в первом разряде и во втором разряде элементов массива. Несколькими способами вычисляются «расстояния» от полученных эмпирических частот до теоретических частот закона Бенфорда. На множестве точек, координатами которых являются вычисленные расстояния, осуществляется кластерный анализ, разбивающий массивы на две группы, характеризующиеся различной степенью «близости» к закону Бенфорда. Результаты кластерного анализа используются для обучения классификатора на основе логистической регрессии, который в дальнейшем применяется для прогнозирования наличия (или отсутствия) искажений в финансовой отчетности, получаемой от новых компаний.

Ключевые слова: закон Бенфорда; фальсификация отчетности; кластерный анализ; логистическая регрессия.

Для цитирования: Криволапов С.Я., Комиссарова А.В., Хамула Д.А. Использование распределения Бенфорда для снижения риска необнаружения искажений финансовой отчетности // Проблемы анализа риска. 2025. Т. 22. № 1. С. 88–95. — EDN: LTXGRW

Авторы заявляют об отсутствии конфликта интересов

Using the Benford Distribution to Reduce the Risk of Undetected Misstatements of Financial Statements

Sergey Y. Krivolapov*,
Anna V. Komissarova,
Daniil A. Khamula,
Financial University under the Government of the Russian Federation,
Leningradsky av., 49, Moscow, 125167, Russia

Abstract
A set of numerical arrays is considered, each of which describes the economic activities of some companies. For each array, the frequencies of occurrence of each of the possible digits in the first digit and in the second digit of the array elements are determined. The "distances" from the obtained empirical frequencies to the theoretical frequencies of Benford's law are calculated in several ways. Cluster analysis is performed on a set of points whose coordinates are calculated distances, dividing arrays into two groups characterized by varying degrees of "proximity" to Benford's law. The results of cluster analysis are used to train a classifier based on logistic regression, which is then used to predict the presence (or absence) of distortions in financial statements received from new companies.

Keywords: Benford's law; falsification of reports; cluster analysis; logistic regression.

For citation: Krivolapov S.Y., Komissarova A.V., Khamula D.A. Using the Benford distribution to reduce the risk of undetected misstatements of financial statements // Issues of Risk Analysis. 2025;22(1):88-95. (In Russ.). — EDN: LTXGRW

The authors declare no conflict of interest

Содержание

Введение

Данные о финансовом состоянии организации важны как для внутренних, так и для внешних пользователей. Внешние пользователи в лице контролирующих органов, банков или инвесторов могут обращаться к бухгалтерским документам для того, чтобы проверить компанию на благонадежность и оценить, как хорошо она исполняет свои обязательства перед государством и собственными работниками.

В большинстве случаев финансовая отчетность компаний подлежит обязательному аудиту. Однако существует высокий риск того, что финансовая отчетность будет содержать искаженные данные, а аудиторская проверка не сможет гарантировать отсутствие фальсификации отчетных показателей. Существует ненулевая вероятность так называемого *аналитического риска* — риска того, что при проверке выбранной совокупности проверяемых объектов используемые аудиторские процедуры не позволят обнаружить имеющиеся ошибки.

Чтобы обеспечить сбор достаточно весомой доказательной базы для вынесения конечного вердикта компании относительно наличия каких-либо финансовых манипуляций в ее деятельности, необходим широкий инструментарий, позволяющий выявлять несоответствия в отчетности фирм. Одним из инструментов для выявления несоответствий между поддельными данными и фактическими является закон Бэнфорда или закон первой цифры [1].

Этот закон показывает вероятность появления определенной цифры в первом разряде величин, описывающих различные процессы жизнедеятельности, в том числе при составлении финансовых отчетов.

Отклонение эмпирической частоты появления цифр в первом разряде числовых данных может являться признаком наличия определенных искажений в этих показателях. Примеры и методы применения закона Бенфорда для выявления искажений в данных можно найти в работах [2–10]. В указанных работах рассматривается применение закона Бенфорда для анализа данных различных предметных областей: проверка финансовой отчетности [1–5], анализ банковских транзакций [6], контроль изображений отпечатков пальцев [7, 8], анализ контрактов [9], проверка результатов голосования [10].

Задача выявления расхождений с законом Бенфорда решается путем оценки статистической значимости соответствия эмпирических частот цифр первого и второго разрядов чисел данных теоретическим частотам закона Бенфорда.

Недостатком такого решения является достаточно большая степень субъективности. Решение принимается на основе задаваемого пользователем уровня значимости (вероятности ошибочного отклонения гипотезы о соответствии распределения первой цифры данных распределению Бенфорда).

В представленной работе предложено решение, использующее числовую меру, характеризующую степень отличия распределения первой цифры от распределения Бенфорда для первой цифры и распределения второй цифры от распределения Бенфорда для второй цифры. В качестве такой меры используется кросс-энтропия — показатель, являющийся частью расстояния Кульбака-Лейблера. Само по себе значение кросс-энтропии не имеет содержательного описания, но его можно использовать для сравнения: если проверяются два массива числовой информации на соответствие их распределения закону Бенфорда, то массив, у которого кросс-энтропия меньше, точнее описывается этим законом.

Предлагается следующий алгоритм действий:

1. Исходные данные.

Берется некоторое количество массивов, содержащих финансовую отчетность различных компаний. Требования, которым должны удовлетворять массивы: они должны иметь различную степень достоверности данных, различную степень имеющихся в них искажений. Чем большее разнообразие по степени достоверности информации будет представлено в совокупности массивов, тем лучше. Оптимальным представляется набор из нескольких десятков компаний.

2. Вычисление признаков, численно характеризующих близость распределения цифр массивов данных к распределению Бенфорда.

Для каждого массива вычисляются два признака: x_1 — кросс-энтропия между распределением первой цифры и распределением Бенфорда для первой цифры; x_2 — кросс-энтропия между распределением второй цифры и распределением Бенфорда для второй цифры.

3. Подготовка обучающего набора данных на основе кластерного анализа.

На множестве точек (x_{1i}, x_{2i}), (i = 1, 2, …, n; n — число компаний) проводится кластерный анализ, позволяющий разбить это множество на некоторое

количество кластеров (сравнительно однородных групп). Полученные кластеры можно дифференцировать по степени близости признаков (x_{1i}, x_{2i}) к нулю.

4. Предсказание с использованием классификатора.

Имея обучающий набор данных, можно для нового проверочного массива делать прогноз: к какому из кластеров обучающего набора относится этот массив.

Классификатор (это может быть наивный байесовский классификатор, логистическая регрессия или метод *k* ближайших соседей) определяет вероятности принадлежности проверочного массива данных тому или иному кластеру. В качестве предсказания подходящего кластера указывается тот, для которого была получена максимальная из вычисленных вероятностей.

В случае, когда кластерный анализ проводился с числом кластеров — два, предсказание кластера с наименьшими значениями признаков (x_{1i}, x_{2i}) можно трактовать как наличие массива с достоверными данными. В случае числа кластеров больше двух можно расширить варианты ответов прогноза («массив с достоверными данными», «массив с недостоверными данными», «массив с промежуточными (неопределенными) данными»).

1. Закон Бенфорда

Пусть генеральная совокупность *X* состоит из чисел, записанных в десятичной системе счисления и имеющих, по крайней мере, два разряда ненулевой целой части. Обозначим через *Y* случайную величину, равную цифре, стоящей в первом разряде числа совокупности *X*. Закон Бенфорда [1] утверждает, что для совокупности *X*, основанной на данных источников реальной жизни, функция вероятности случайной величины *Y* имеет вид:

$$p_j = \mathbf{P}(Y = j) = \lg(j + 1) - \lg(j), j = 1, 2, \ldots, 9.$$

Продемонстрируем выполнение закона Бенфорда для некоторых реальных данных. В массив *X* записаны данные о размере площади стран мира, вслед за которыми записаны данные о населении стран мира. На рис. 1 приведен график функции вероятности закона Бенфорда и эмпирические вероятности цифр в первом разряде чисел массива *X*.

Закон Бенфорда допускает обобщения на цифры, составленные из разрядов чисел, начиная со второго и далее[1] [11].

[1] https://en.wikipedia.org/wiki/Benford%27s_law

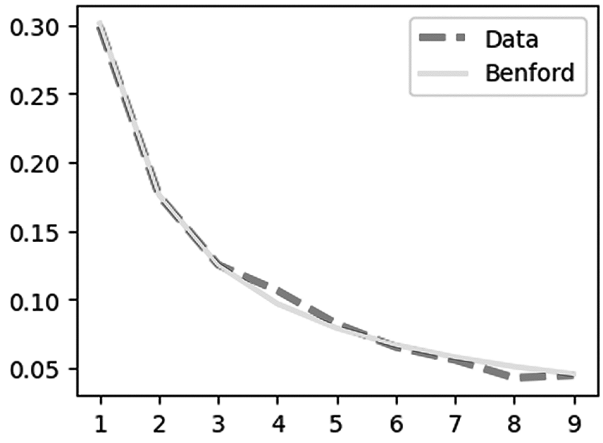

Рис. 1. График теоретических и эмпирических вероятностей первой цифры

Figure 1. Graph of theoretical and empirical probabilities of the first digit

Пусть совокупность *X* содержит числа, записанные в десятичной системе счисления и с целой частью, содержащей, по крайней мере, два разряда. Пусть *Z* — случайная величина, равная двузначному числу, составленному из первых двух разрядов. Обобщение закона Бенфорда для случайной величины *Z* гласит, что функция вероятности величины *Z* имеет вид:

$$q_l = \mathbf{P}(Z = l) = \lg(l + 1) - \lg(l), l = 10, 11, \ldots, 99.$$

Предполагая справедливым закон Бенфорда для первой цифры и для первого двузначного числа, можно получить вид функции вероятности для случайной величины *U*, равной отдельной второй цифре числа:

$$\mathbf{P}(U = j) = \sum_{k=1}^{9} \lg\left(1 + \frac{1}{10k + j}\right).$$

2. «Расстояние» между распределениями

Классическим способом оценки близости распределений является расстояние Кульбака-Лейблера. Для дискретных распределений f_k и g_k вычисляется по формуле:

$$D_{KL} = \sum_k f_k \ln \frac{f_k}{g_k}.$$

В нашем случае распределением *f* является закон Бенфорда, распределение *g* — эмпирические распределения, полученные из массивов данных компаний. Для нас важным является сравнительное значение

«расстояний» для различных компаний. Если записать формулу для расстояния Кульбака-Лейблера в виде:

$$D_{KL} = \sum_k f_k \ln \frac{f_k}{g_k} = \sum_k f_k \ln f_k - \sum_k f_k \ln g_k,$$

можно сделать вывод, что для сравнения «расстояний» между распределением f и различными распределениями g достаточно использовать только второе слагаемое, так как первое слагаемое от распределения g не зависит.

Соответствующее выражение:

$$D = -\sum_k f_k \ln g_k$$

имеет наименование «кросс-энтропия» и используется в различных программах машинного обучения.

3. Предсказание на основе работы классификатора

В представленной работе рассматривается вариант разбиения данных на две группы (с числом кластеров, равным двум; массивы с достоверными данными и массивы с недостоверными данными). С учетом этого факта для целей предсказания была выбрана бинарная логистическая регрессия.

Бинарную классификацию можно визуализировать. Наши два кластера образуют два облака, отделенных друг от друга (их условно называют класс «+» и класс «–»). Между облаками проводится разделяющая их плоскость (линейный дискриминант с уравнением $y = ax_1 + bx_2 + c$). При подстановке в уравнение дискриминанта координат точки, отвечающей нашему проверочному массиву, получится некоторое число t. Если оно будет положительным, точка находится со стороны класса «+», она будет отнесена к кластеру 1; в случае $t < 0$ точка находится со стороны класса «–» и она будет отнесена к кластеру 2. Для получения вероятности отнесения точки к определенному кластеру используется логистическая функция. Вероятность попадания точки в класс «+» вычисляется по формуле:

$$P_+ = \frac{e^t}{e^t + 1}.$$

Вероятность попадания в класс «–»: $P_- = 1 - P_+$.

4. Объем данных компаний

Проверка на соответствие эмпирических частот теоретическим частотам закона Бенфорда предполагает использование большого массива данных.

При исследовании вопроса о необходимом размере массивов использовался следующий факт: закону распределения Бенфорда очень хорошо соответствуют объемы продаж акций российских компаний (их соответствие вполне сопоставимо с соответствием таких классических примеров для закона Бенфорда, как численность населения и размер площадей стран мира). На рис. 2 приведен график функции вероятности закона Бенфорда и эмпирические вероятности цифр в первом разряде чисел массивов: «Численность населения стран мира», «Объемы продаж акций компании «Башнефть», «Объемы продаж акций компании «Лензолото».

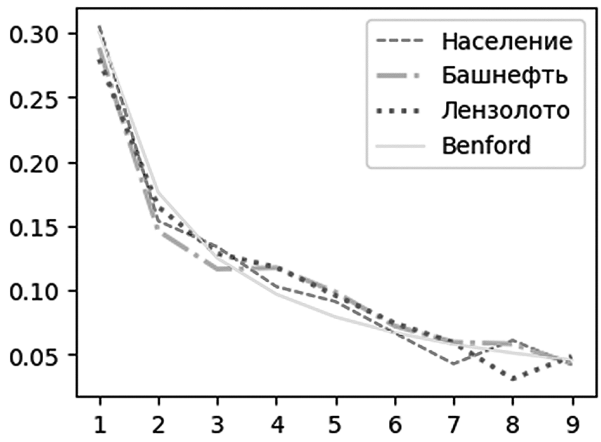

Рис. 2. Эмпирические вероятности первой цифры некоторых компаний

Figure 2. Empirical probabilities of the first digit of some companies

Данные, касающиеся объемов продаж акций, доступны в любых требуемых объемах[2] [12]. При увеличении длины массива данных степень соответствия закону распределения Бенфорда увеличивается. На рис. 3 приведены эмпирические вероятности первой цифры для объема продаж акций компании «Башнефть» длиной 500 и 5000.

Вычислялись значения «расстояний» (признаков x_1 и x_2) до теоретического закона распределения для массивов, содержащих объемы продаж акций различной длины (500, 1000, 2000, 3000, 4000 и 5000). Получены следующие результаты:

[[4.5138847 4.7797392]
[3.81286855 4.08236587]
[3.12204432 3.38687422]

[2] https://mfd.ru/export/

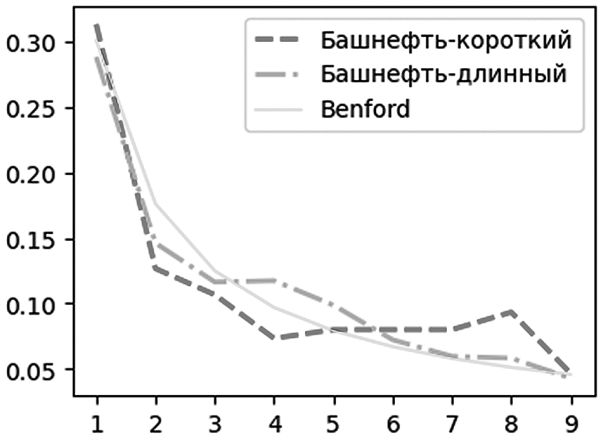

Рис. 3. Эмпирические вероятности первой цифры для данных различной длины

Figure 3. Empirical probabilities of the first digit for data of different lengths

[2.71353516 2.97866075]
[2.42378512 2.71024354]
[2.38100096 2.67878585]]

Близкие результаты были получены для данных различных компаний. Для объемов порядка $n = 5000$ происходит стабилизация значений признаков. Все дальнейшие расчеты проводились на массивах данных объема 5000.

Демонстрация описанной процедуры на данных, содержащих сведения по 36 компаниям. Расчеты проводились на языках Python и R.

1. Чтение данных.

Результат в таблице датафрейм с 36 столбцами.

2. Вычисление эмпирических вероятностей.

Для каждого столбца (для каждой компании) вычисляются эмпирические вероятности (относительные частоты) $\hat{p}_{ij}$ — появления определенной цифры в первом разряде элементов столбца и $\hat{q}_{il}$ — появления определенной цифры во втором разряде числа, $i = 1, 2, \ldots, 36$; $j = 1, 2, \ldots, 9$; $l = 0, 1, \ldots, 9$.

Некоторые цифры могут не встречаться в исходных данных. Чтобы исключить появление нулевых значений частот (в дальнейшем от них будет браться логарифм), нулевое значение заменялось малым положительным числом (10^{-5}).

3. Формирование признаков x_{i1}, x_{i2}.

Для каждого из 36 столбцов вычисляются «расстояния» (кросс-энтропия, формула (1)) x_{i1}, x_{i2} между теоретическими вероятностями закона Бенфорда p_{ij}, q_{il} и эмпирическими вероятностями $\hat{p}_{ij}$, $\hat{q}_{il}$, $i = 1, 2, \ldots, 36$; $j = 1, 2, \ldots, 9$; $l = 0, 1, \ldots, 9$.

$$x_{i1} = -\sum_{j=1}^{9} p_{ij} \ln \hat{p}_{ij},\ x_{i2} = -\sum_{l=0}^{9} q_{il} \ln \hat{q}_{il}.$$

4. Нормирование признаков.

Для приведения к сопоставимому виду признаки нормировались согласно формуле $(x_i - x_{\min})/(x_{\max} - x_{\min})$.

Из нормированных значений признаков формировалась таблица датафрейм.

5. Кластерный анализ.

На множестве точек (x_{i1}, x_{i2}), $i = 1, 2, \ldots, 36$ проводился кластерный анализ методом k-средних с числом кластеров $k = 2$. Визуализация результата приведена на рис. 4.

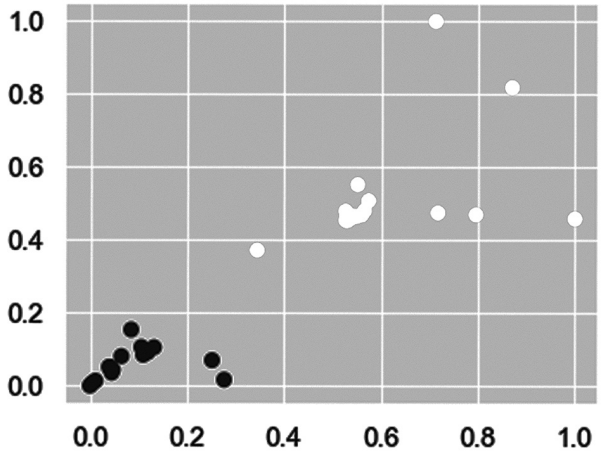

Рис. 4. Результат кластерного анализа с числом кластеров 2

Figure 4. The result of cluster analysis with the number of clusters 2

Метки кластеров:

[0010001111111010101000001111110101011]

Точки кластера 0 соответствуют компаниям с малыми значениями признаков-расстояний x_1, x_2.

Результат работы — размеченная выборка: матрица признаков X размера 36×2 и массив меток y длиной 36 (значения меток: 1 — недостоверная информация, 0 — достоверная информация).

6. Выполнение предсказания.

Для выполнения предсказания о степени достоверности новых данных использовалась логистическая регрессия (функция LogisticRegression() пакета sklearn.linear_model языка Python).

Размеченные данные (*X*, *y*) делились на две части: тренировочный набор (*Xtrain*, *ytrain*) и тестовый набор (*Xtest*, *ytest*). Тестовый набор формировался путем случайного выбора шести значений из размеченного набора (*X*, *y*). В тренировочный набор вошли данные оставшихся тридцати компаний.

В целях контроля точности будущего предсказания меток тестового набора запишем, какие они имели метки после кластерного анализа.

Результат случайного выбора номеров компаний тестового набора: 3, 16, 19, 20, 29, 36. Выше был приведен массив меток. На указанных позициях стоят метки: 1, 0, 1, 0, 1, 1.

На вход классификатора подавалась размеченная тренировочная выборка (*Xtrain*, *ytrain*).

```
from sklearn.linear_model import \
LogisticRegression
lr = LogisticRegression()
lr_model = lr.fit(Xtrain, ytrain)
```

По обучающей выборке классификатор строит разделяющую плоскость.

Теперь, подавая на вход функции lr.predict() набор признаков *Xtest* тестового набора, можно получить предсказание: классификатор оценит вероятность попадания точки (x_1, x_2) по одну или другую сторону разделяющей плоскости.

```
lr_predictions = lr.predict(Xtest)
```

Результат предсказания:

```
print(lr_predictions)
[1,0,1,0,1,1]
```

Все метки тестового набора предсказаны точно.

Проверим работоспособность предложенной процедуры, выполнив предсказание еще для двух массивов: первый содержит раздел финансовой отчетности компании «Роснефть»; второй — раздел отчетности компании «Трансаэро» (компании, у которой были вскрыты проблемы с отчетностью).

Вычисляем для одного и второго массивов эмпирические вероятности $\hat{p}_{ij}, \hat{q}_{il}$, $i = 1, 2$; $j = 1, \ldots, 9$; $l = 0, \ldots, 9$ появления первой цифры и второй цифры в исходных данных. Формируем признаки («расстояния») (x_{i1}, x_{i2}), $i = 1, 2$ и записываем их в матрицу *Xtest*. Вызываем функцию lr.predict().

```
lr_predictions = lr.predict(Xtest)
```

Результат предсказания:

```
print(lr_predictions)
```

[0,1]

Для компании «Роснефть» результат: «Массив данных содержит достоверную информацию», для компании «Трансаэро»: «Массив данных содержит недостоверную информацию».

Логистический классификатор позволяет посмотреть, с какой вероятностью производится назначение определенной метки класса.

```
prob_test = lr.predict_proba(Xtest)
print(prob_test)
[[0.0528321 0.9471679]
[0.8993031 0.1006969]]
```

В рассмотренном примере метка класса 0 (для компании «Роснефть») предсказана с вероятностью 0,947; метка класса 1 (для компании «Трансаэро») — с вероятностью 0,899.

Заключение

В работе изложена процедура обучения и использования для предсказания классификатора, основанного на проверке близости частот цифры первого разряда и цифры второго разряда теоретическим частотам закона Бенфорда.

Для обучения классификатора был использован кластерный анализ методом *k*-средних с числом кластеров 2.

Вычисления, проведенные для реальных данных, содержащих сведения о финансовой отчетности компаний, показали работоспособность предложенной процедуры.

Список источников [References]

1. Кечкова И. В., Кеворкова Ж. А. Закон Бенфорда как метод выявления мошеннических действий // Молодой ученый. 2017. № 11(145). С. 219–221 [Kechkova I. V., Kevorkova J. A. Benford's law as a method of detecting fraudulent actions // Young Scientist. 2017;(11(145)).:219–221. (In Russ.)]
2. Назарова В. В., Чуракова И. Ю., Куприянов В. А. Проверка достоверности финансовой отчетности европейских компаний законом Бенфорда // AlterEconomics. 2023. Т. 20. № 3. С. 691–711. https://doi.org/10.31063/AlterEconomics/2023.20–3.10 [Nazarova V. V., Churakova I. Yu., Kupriyanov V. A. Assessing financial statement reliability in european companies using Benford's law // AlterEconomics. 2023;20(3):691–711. (In Russ.). https://doi.org/10.31063/AlterEconomics/2023.20–3.10]

3. Зверев Е., Никифоров А. Распределение Бенфорда: выявление нестандартных элементов в больших совокупностях финансовой информации // Внутренний контроль в кредитной организации. 2018. № 4. С. 4–18 [Zverev E., Nikiforov A. Benford distribution: identification of non-standard elements in large sets of financial information // Internal control in a credit institution. 2018;(4):4–18. (In Russ.)]
4. Herteliu, C., Jianu, I., Dragan, I.M., Apostu, S. and Luchian, I. (2021). Testing Benford's Laws (non)conformity within disclosed companies' financial statements among hospitality industry in Romania. Physica A: Statistical Mechanics and its Applications. 582p. 126221 https://doi.org/10.1016/j.physa.2021.126221
5. Durtschi, Cindy & Hillison, William & Pacini, Carl. (2004). The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. J. Forensic Account. 5
6. Žgela, Mario & Krakar, Zdravko. (2009). Application of Benford's Law in Payment Systems Auditing. Journal of Information and Organizational Sciences. 33
7. Nigrini, M. (1996) A Taxpayer Compliance Application of Benford's Law. The Journal of the American Taxation Association, 18, 72–91
8. Nigrini, Mark J. and Linda Jean Mittermaier. "The Use of Benford's Law as an Aid in Analytical Procedures." Auditing-a Journal of Practice & Theory 16 (1997):52
9. Silva, Wilton & Travassos, Silvana & Costa, José. (2017). Using the Newcomb-Benford Law as a Deviation Identification Method in Continuous Auditing Environments: A Proposal for Detecting Deviations over Time. Revista Contabilidade & Financas. 28. 11–26 https://doi.org/10.1590/1808-057x201702690
10. Pericchi, Luis & Torres, David. (2012). Quick Anomaly Detection by the Newcomb — Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela. Statistical Science — STAT SCI. 26. 10.1214/09-STS296

Сведения об авторах

Криволапов Сергей Яковлевич: кандидат физико-математических наук, доцент, доцент Финансового Университета при Правительстве Российской Федерации
Количество публикаций: 90, в т.ч. 10 учебников
Область научных интересов: теория вероятностей, математическая статистика, анализ данных
Scopus Author ID: MFZ-7354-2025
ORCID: 0009-0009-4745-8047
SPIN-код: 7149-9620
Контактная информация:
Адрес: 125167, г. Москва, Ленинградский пр-т, д. 49
skrivolapov@fa.ru

Комиссарова Анна Владимировна: студент Факультета экономики и бизнеса Финансового Университета при Правительстве Российской Федерации
ORCID: 0009-0005-3952-4631
SPIN-код: 3086-9234
Контактная информация:
Адрес: 125167, г. Москва, Ленинградский пр-т, д. 49
annakomissarova04@gmail.com

Хамула Даниил Александрович: студент Факультета экономики и бизнеса Финансового Университета при Правительстве Российской Федерации
ORCID: 0000-0001-9633-3747
Контактная информация:
Адрес: 125167, г. Москва, Ленинградский пр-т, д. 49
khamula.2003@mail.ru

Статья поступила в редакцию: 06.06.2024
Одобрена после рецензирования: 07.10.2024
Принята к публикации: 25.10.2024
Дата публикации: 28.02.2025

The article was submitted: 06.06.2024
Approved after reviewing: 07.10.2024
Accepted for publication: 25.10.2024
Date of publication: 28.02.2025