**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ**
**И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 1 С. 197–201**

Ки&М

**МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ**

УДК: 51-78, 519.234.3, 519.257, 81-139

# Отклонения от закона Бенфорда и распознавание авторских особенностей в текстах

**А. В. Зенков**

Уральский федеральный университет,
Россия, 620002, г. Екатеринбург, ул. Мира, д. 19

E-mail: zenkow@mail.ru

*Получено 05 октября 2014 г.*

Исследовано распределение первой значащей цифры в числительных связных текстов. Обнаружено, что закон Бенфорда приближенно выполняется для них. Отклонения от закона Бенфорда являются статистически устойчивыми авторскими особенностями, позволяющими при некоторых условиях различить части текста с разным авторством.

Ключевые слова: закон Бенфорда, статистическая проверка гипотез, критерий Манна–Уитни

# Deviation from Benford’s law and identification of author peculiarities in texts

**A. V. Zenkov**

*Ural federal university, 19 Mira st., Ekaterinburg, 620002, Russia*

**Abstract.** — The distribution of the first significant digit in numerals of connected texts is considered. Benford's law is found to hold approximately for them. Deviations from Benford's law are statistically significant author peculiarities that allow, under certain conditions, to distinguish between parts of the text with a different authorship.

Keywords: Benford’s law, Statistical hypothesis testing, Mann–Whitney U-test

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 1, pp. 197–201 (Russian).

## Введение

Своеобразное проявление закона больших чисел — известный уже более ста лет закон Бенфорда [Benford, 1938] — в последние десятилетия из статистического курьеза превращается в полезное средство анализа данных. Закон Бенфорда описывает вероятность появления определенной первой значащей цифры в разнообразных распределениях величин, взятых из реальной жизни. Вопреки здравому предположению о том, что частоты появления любой первой значащей цифры должны быть равными, для многих массивов данных в качестве первой значащей цифры чаще других встречается единица! Согласно закону Бенфорда вероятность $P(d)$ появления цифры $d$ в качестве первой значащей

$$P(d) = \lg\left(1 + \tfrac{1}{d}\right), \tag{1}$$

так что $d = 1$ должна встречаться с вероятностью $\lg 2 \approx 0.30$, $d = 2$ — с вероятностью 0,18 и т. д.

Исчерпывающего объяснения закона Бенфорда, охватывающего все случаи реализации, до сих пор не предложено, хотя и сформулированы некоторые условия, благоприятствующие его появлению. Один из классических опытов Бенфорда, хорошо согласующийся с (1), — анализ встречаемости числительных в статьях случайно выбранного выпуска популярного журнала — находит логичное объяснение в теореме [Hill, 1995], согласно которой в условиях неоднократного *случайного* выбора распределения вероятностей с последующим *случайным* выбором числа согласно этому распределению возникает набор чисел, подчиняющийся закону Бенфорда. Как мы покажем ниже, эти условия не являются необходимыми: даже для *связного* текста, к которому условия теоремы неприменимы, наблюдается распределение первых значащих цифр числительных, близкое к (1).

Несмотря на неполноту объяснения, закон Бенфорда успешно применяется для выявления подлогов в бухгалтерской отчетности [Nigrini, 2012] и фальсификаций на выборах [Battersby, 2009]; обсуждаются применения в различных областях от сейсмологии [Sambridge, Tkalčić, Arroucau, 2011] до стеганографии [Andriotis, Oikonomou, Tryfonas, 2013].

Цель настоящей работы — показать, что при определенных условиях исследование частот появления различных значащих цифр в связном тексте может быть полезным с точки зрения вопроса об авторстве текста, поскольку эти частоты специфичны для автора.

## Статистическое исследование текстов

Для поставленных целей наиболее показателен анализ разных произведений одного автора либо произведений разных авторов на близкие темы.

Исследованию были подвергнуты:

1) Записки Юлия Цезаря о Галльской войне и о Гражданской войне;

2) четыре канонических Евангелия — от Матфея, Марка, Луки и Иоанна.

Для этих текстов исследовались частоты появления различных первых значащих цифр с учетом количественных и порядковых числительных, выраженных как цифрами, так и (значительно чаще) словесно.

Известно, что первые семь книг «Записок о Галльской войне» написаны Цезарем, а последняя, восьмая, книга — Гирцием, завершившим произведение. На рисунках 1–4 приведены распределения частот встречаемости первой значащей цифры, характерные для Цезаря и Гирция. При общем приближенном выполнении закона Бенфорда (1) легко заметны характерные авторские различия. Для всего произведения характерно более редкое появление единицы в качестве первой значащей цифры по сравнению с предписанием закона Бенфорда и обратное явление для цифры 2 (рис. 1). Это авторская особенность стиля Цезаря: она проявляется в принадлежащих его перу книгах «Записок» (на рис. 2, 3 представлены для сравнения результаты

для первой и третьей книг «Записок»). Совершенно иной вид имеет аналогичное распределение для восьмой книги «Записок», автором которой является Гирций (рис. 4).

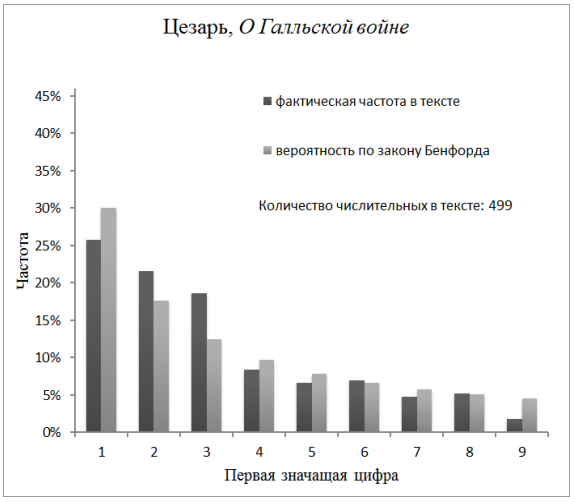


Рис. 1. Распределение первых значащих цифр числительных в «Записках о Галльской войне» Цезаря. Результаты обработки всех 8 книг «Записок»

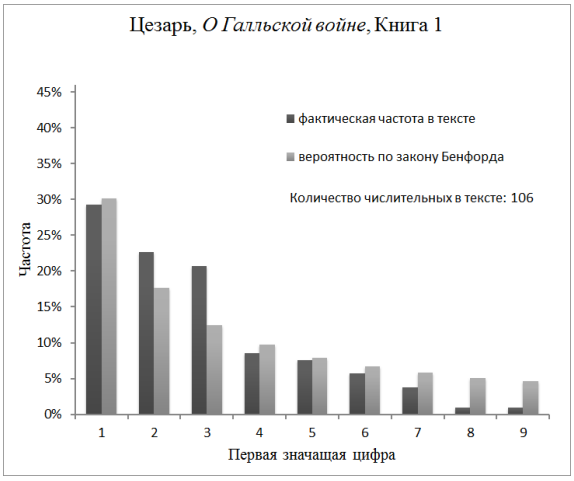


Рис. 2. Распределение первых значащих цифр числительных в «Записках о Галльской войне» Цезаря. Результаты обработки только 1-й книги «Записок», написанной Цезарем


Рис. 3. Распределение первых значащих цифр числительных в «Записках о Галльской войне» Цезаря. Результаты обработки только 3-й книги «Записок», написанной Цезарем

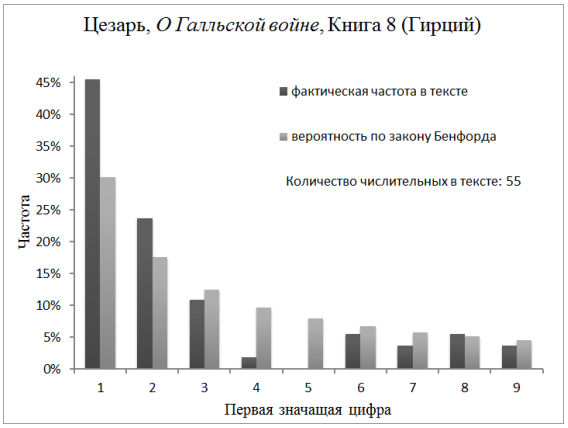


Рис. 4. Распределение первых значащих цифр числительных в «Записках о Галльской войне» Цезаря. Результаты обработки только 8-й книги «Записок», написанной Гирцием

В другом произведении Цезаря — «Записках о Гражданской войне», написанных им единолично, — наблюдается та же характерная особенность (рис. 5). Поскольку это произведение посвящено иным историческим событиям, данное совпадение следует расценивать как отражение именно *стиля* Цезаря.

В качестве еще одного примера произведений разных авторов с общей тематикой нами рассмотрены Евангелия от Матфея, Марка, Луки и Иоанна. В целом распределение первой значащей цифры числительных и здесь напоминает распределение Бенфорда, но заметно преобладает единица (рис. 6–9). Между распределениями (особенно первыми тремя и последним) наблюдаются различия — не очень большие, но статистически значимые, с учетом количества проанализированных данных.

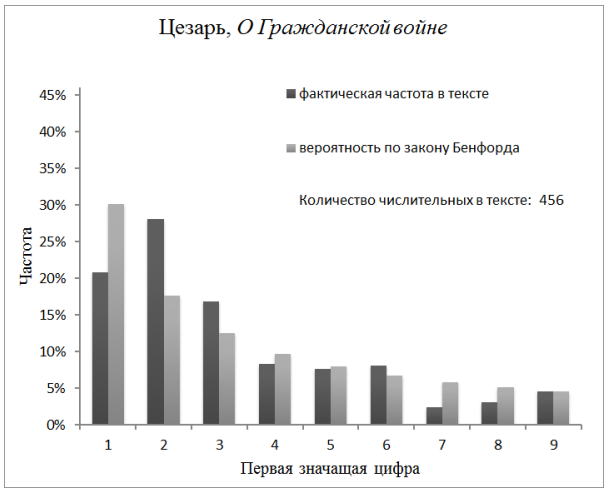


Рис. 5. Распределение первых значащих цифр числительных в «Записках о Гражданской войне» Цезаря

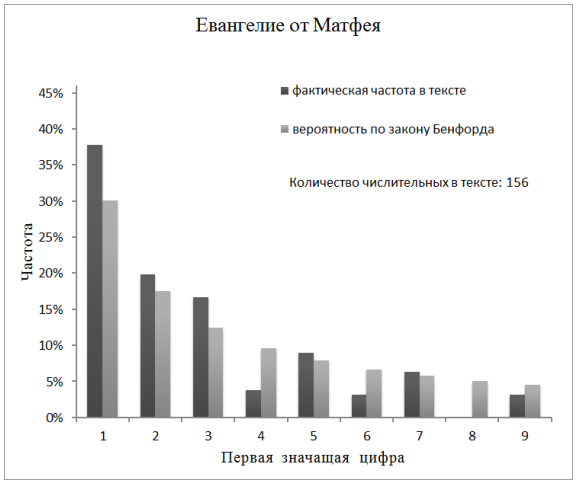


Рис. 6. Распределение первых значащих цифр числительных в Евангелии от Матфея

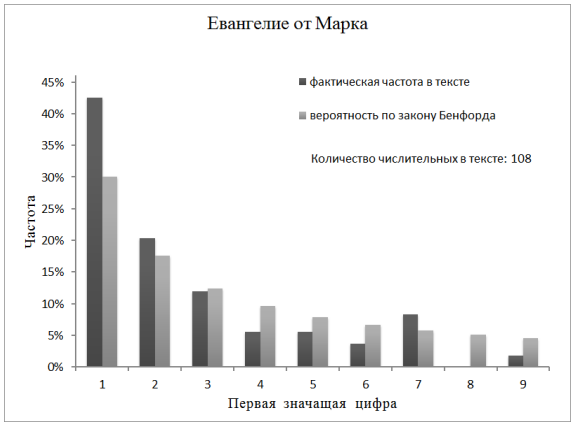


Рис. 7. Распределение первых значащих цифр числительных в Евангелии от Марка

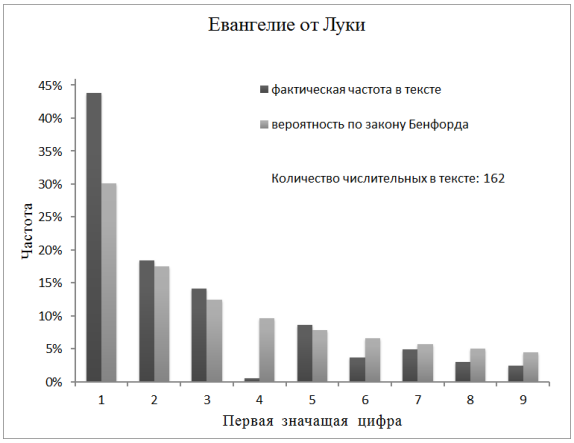


Рис. 8. Распределение первых значащих цифр числительных в Евангелии от Луки

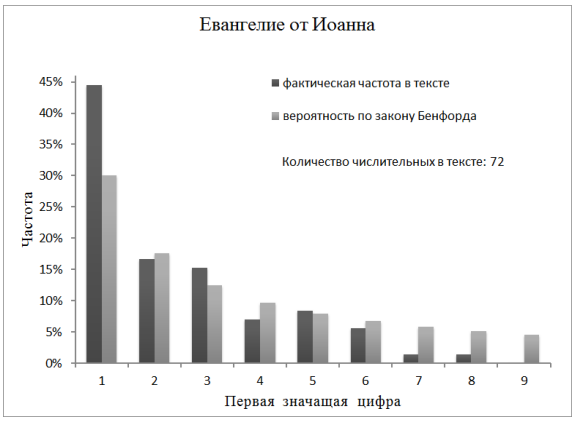


Рис. 9. Распределение первых значащих цифр числительных в Евангелии от Иоанна

Известное автору исследование [Hüngerbühler, 2007] распределения первых значащих цифр числительных в корпусе Священного Писания преследовало иные цели.

Разумеется, сравнение распределений не может основываться только на выявлении субъективных визуальных сходства и различий между ними. Нами применен непараметрический U-критерий Манна–Уитни. Нулевая гипотеза $H_0$, утверждающая *отсутствие* значимых различий в рассмотренных распределениях, оказалась отвергнутой и принятой именно в тех случаях, как описано выше. А именно, различие между книгами «записок о Галльской войне», написанными Цезарем и Гирцием, оказалось значимым $(p = 0,02)$, а между разными книгами Цезаря — не значимыми $(p = 0,25)$. Различия между Евангелиями от Матфея, Марка, Луки, с одной стороны, и Евангелием от Иоанна, с другой, оказались значимыми $(p = 0,03 \div 0,04)$, а между любыми двумя Евангелиями из первой тройки — *не значимыми*. Итак, предлагаемый нами метод разграничения авторства не всесилен, но может быть полезным дополнением к традиционным методам [Manning, Schütze, 1999].

## Заключение

Закон Бенфорда приближенно выполняется для связных текстов.

Отклонения от закона Бенфорда являются статистически значимыми авторскими особенностями, позволяющими при некоторых условиях различить части текста с разным авторством. Очевидными требованиями является достаточная длина текста и употребительность числительных в нем, чему, например, как правило, удовлетворяет историческая литература.

Распределение цифр конца ряда $\{1,2,\ldots,7,8,9\}$ подвержено сильным флуктуациям и непоказательно.

## Список литературы

*Andriotis P., Oikonomou G., Tryfonas T.* JPEG steganography detection with Benford's Law // Digital Investigation. — 2013. — Vol. 9, No. 3–4. — P. 246–257.

*Battersby S.* Statistics hint at fraud in Iranian election // New Scientist. — 24 June 2009.

*Benford F.* The law of anomalous numbers // Proceedings of American Philosophical Society. — 1938. — Vol. 78, No. 4. — P. 551–572.

*Hill T. P.* A Statistical Derivation of the Significant-Digit Law // Statistical Science. — 1995. — Vol. 10 — P. 354–363.

*Hüngerbühler N.* Benfords Gesetz über führende Ziffern: wie die Mathematik Steuersündern das Fürchten lehrt // EducETH, Publikation der Eidgenössischen Technischen Hochschule Zürich. — 2007. — URL: www.educ.ethz.ch/unt/um/mathe/ana/benford

*Manning C. D., Schütze H.* Foundations of Statistical Natural Language Processing. — Cambridge (Mass.) — London: The MIT Press, 1999. — XXXVII + 680 p.

*Nigrini M. J.* Benford's Law: applications for forensic accounting, auditing, and fraud detection. — Hoboken: John Wiley & Sons, Inc., 2012. — XX + 330 pp.

*Sambridge M., Tkalčić H., Arroucau P.* Benford's Law of First Digits: from Mathematical Curiosity to Change Detector // Asia Pacific Mathematics Newsletter. — 2011. — Vol. 1, No. 4. — P. 1–6.