DOI 10.32370/2018_09_3

Benford's Law in Information Science

Yuri N. Klimov, Oleg N. Klimov
Moscow, Russia
E-mail: yuri.klimov.29@mail.ru

Abstract

For the first time, Benford's law is confirmed for computer science and refined by simple algebraic equations for dynamics and cumulative number of digits. For the first time it is also shown by simple algebraic equations the difference of dynamics and cumulates for the relative and relative exponential rates of change F(n) or the probability to meet the figure of the first. Benford's law on lexicology is closer to the quantitative characteristics we have studied in the field of Informatics and other fields of knowledge about the Universe, i.e. it is a universal law.

Keywords: Benford's law, information's science, mathematical modeling

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 [1 *http://baguzin.ru/wp/zakon-benforda-ili-zakon-pervoj-tsifry/*].

Первая цифра	*Вероятность*
1	30,1 %
2	17,6 %
3	12,5 %
4	9,7 %
5	7,9 %
6	6,7 %
7	5,8 %
8	5,1 %
9	4,6 %

Рис. 1. Вероятность встретить первую цифру в данных из источников 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 и чуть менее «красиво» факториалы (см. лист «рис. 2» Excel-файла). 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. 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).** Объём данных должен быть достаточен для применения статистических методов.

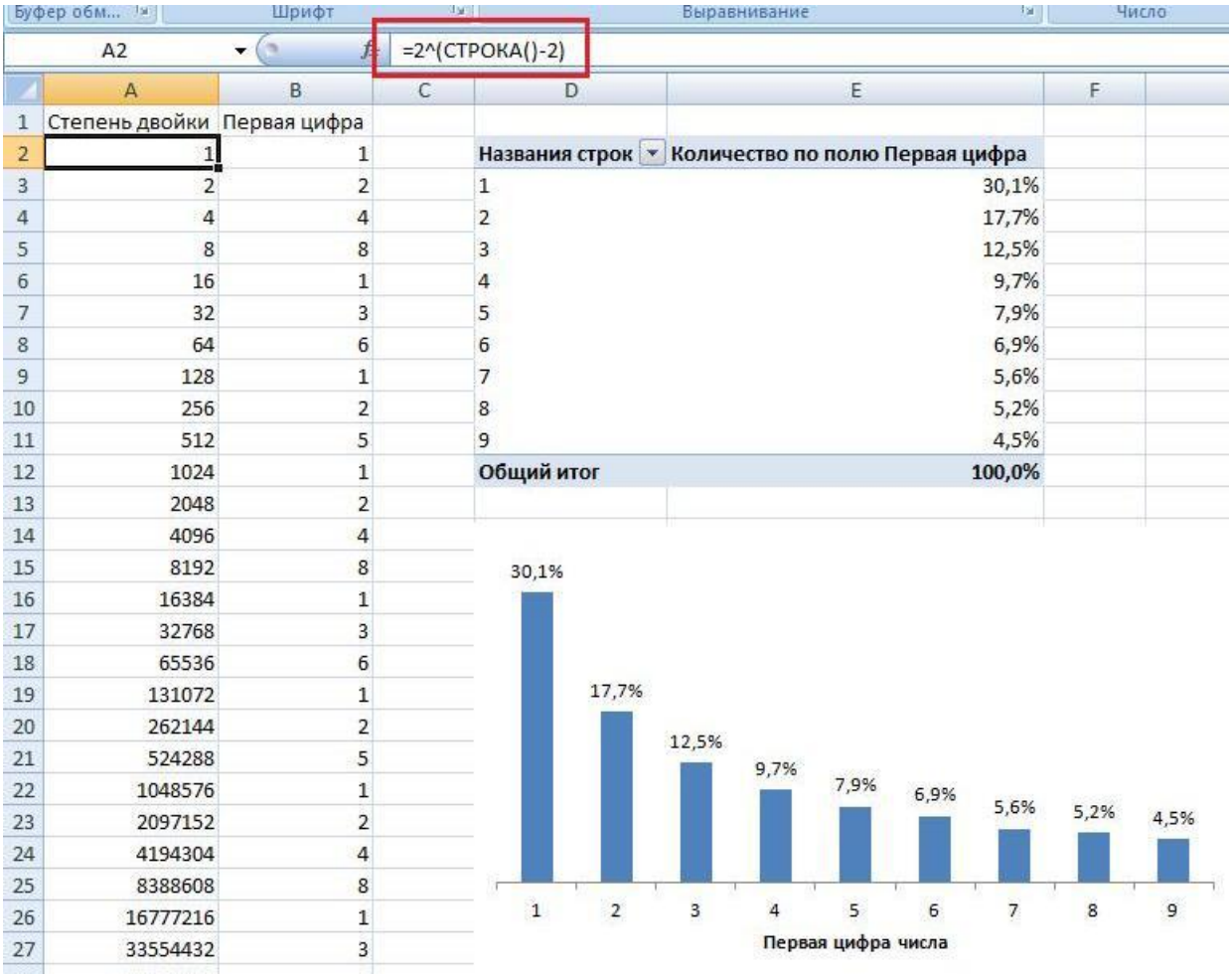

Рис. 2. Первая цифра в числах, являющихся степенью двойки, для диапазона от 2^0 до 2^{1000}

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 (Б Мандельброт. (Не) послушные рынки: фрактальная революция в финансах). Для того чтобы установить явный вид функции *F(n)*, удовлетворяющей закону Бенфорда, рассмотрим переменную величину *G(t),* растущую по показательному (экспоненциальному) закону [2]. 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).

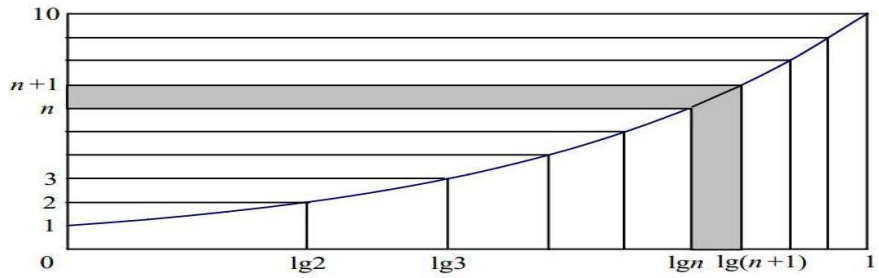

Рис. 3. Объяснение закона Бенфорда

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:

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

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%

Предыдущее описание приведено по работе [1 *http://baguzin.ru/wp/zakon-benforda-ili-zakon-pervoj-tsifry/*].

Выдвигается гипотеза о применимости закона Бенфорда для уточнения вероятности встретить первую цифру в данных и описания простыми алгебраическими уравнениями.

Приведем наши численные данные в области информатики [2]:

- Потоки научно-технической информации;
- Динамика публикаций в Chemical Abstracts (1907-2003)
- Динамика патентов в Chemical Abstracts (1907-2003) ;
- Динамика книг в Chemical Abstracts (1907-2003) ;
- Динамика рефератов в Chemical Abstracts (1907-2003) ;
- Динамика рефератов в РЖ ВИНИТИ РАН ,«Химия» ;
- Динамика английских слогов по Ципфу [6]: ;
- Динамика латинских слогов у Плавта по Ципфу, 1935. [6]: .

Значение *F(n)* или вероятность встретить цифру первой, вычисленные по этой формуле, приведены в табл. 1-8.

Таблица 1. Информатика, Потоки

Информатика, Потоки Первая цифра	Динамика чисел	Значение *F(n)*, или вероятность встретить цифру первой	Кумулятивные числа	Значение *F(n)*, или вероятность встретить цифру первой
1	2	30,103%	2	30,103%
2	1	17,609%	3	47,712%
3	0	12,494%	3	60,206%
4	0	9,691%	3	69,897%

5	1	7,918%	4	77,815%
6	1	6,695%	5	84,510%
7	1	5,799%	6	90,309%
8	1	5,115%	7	95,424%
9	2	4,576%	9	100,000%
	9	100,000%		

Таблица 2. Динамика публикаций в Chemical Abstracts 1907-2003)

Динамика публикаций в CAS (1907-2003) Первая цифра	Динамика чисел	Значение *F(n)*, или вероятность встретить цифру первой	Кумулятивные числа	Значение *F(n)*, или вероятность встретить цифру первой
1	23	30,103%	23	30,103%
2	15	17,609%	38	47,712%
3	26	12,494%	64	60,206%
4	11	9,691%	75	69,897%
5	9	7,918%	84	77,815%
6	5	6,695%	89	84,510%
7	3	5,799%	92	90,309%
8	1	5,115%	93	95,424%
9	3	4,576%	96	100,000%
	96	100,000%		

Таблица 3. Динамика патентов в Chemical Abstracts (1907-2003)

Динамика патентов в CAS (1907-2003) Первая цифра	Динамика чисел	Значение *F(n)*, или вероятность встретить цифру первой	Кумулятивные числа	Значение *F(n)*, или вероятность встретить цифру первой
1	29	30,103%	29	30,10%
2	13	17,609%	42	47,712%
3	9	12,494%	51	60,206%
4	8	9,691%	59	69,897%
5	10	7,918%	69	77,815%
6	7	6,695%	76	84,510%
7	9	5,799%	85	90,309%
8	6	5,115%	91	95,424%
9	6	4,576%	97	100,000%
	97	100,000%		

Таблица 4. Динамика книг в Chemical Abstracts (1907-2003)

Динамика книг в CAS (19 7-2003) Первая цифра	Динамика чисел	Значение *F(n)*, или вероятность встретить цифру первой	Кумулятивные числа	Значение *F(n)*, или вероятность встретить цифру первой
1	33	30,103%	33	30,103%
2	11	17,609%	44	47,712%
3	14	12,494%	58	60,206%
4	8	9,691%	66	69,897%
5	13	7,918%	79	77,815%
6	6	6,695%	85	84,510%

7	5	5,799%	90	90,309%
8	3	5,115%	93	95,424%
	93	100,000%		

Таблица 5. Динамика рефератов в Chemical Abstracts (1907-2003)

Динамика рефератов в CAS (1907-2003) Первая цифра	Динамика чисел	Значение *F(n)* или вероятность встретить цифру первой	Кумулятивные числа	Значение *F(n)*, или вероятность встретить цифру первой
1	19	30,103%	19	30,10%
2	14	17,609%	33	47,712%
3	12	12,494%	45	60,206%
4	19	9,691%	64	69,897%
5	10	7,918%	74	77,815%
6	11	6,695%	85	84,510%
7	7	5,799%	92	90,309%
8	4	5,115%	96	95,424%
9	1	4,576%	97	100,000%
	97	100,000%		

Таблица 6. Динамика рефератов в РЖ «Химия» ВИНИТИ РАН

Динамика реферато в в РЖ «Химия» Первая цифра	Динамика. чисел	Значение *F(n)* или вероятность встретить цифру первой	Кумулятивные числа	Значение *F(n)* или вероятность встретить цифру первой
1	40	30,103%	40	30,10%
2	4	17,609%	44	47,712%

3	0	12,494%	44	60,206%
4	1	9,691%	45	69,897%
5	1	7,918%	46	77,815%
6	0	6,695%	46	84,510%
7	1	5,799%	47	90,309%
8	3	5,115%	50	95,424%
9	1	4,576%	51	100,000%
	51	100,000%		

Таблица 7. Динамика английских слогов по Ципфу

Динамика английских слогов по Ципфу Первая цифра	Динамика чисел	Значение *F(n)*, или вероятность встретить цифру первой	Кумулятивные числа	Значение *F(n)*, или вероятность встретить цифру первой
1	5	30,103%	5	30,10%
2	1	17,609%	6	47,712%
3	2	12,494%	8	60,206%
4	1	9,691%	9	69,897%
5	4	7,918%	13	77,815%
6	1	6,695%	14	84,510%
7	0	5,799%	14	90,309%
8	1	5,115%	15	95,424%
9	0	4,576%	15	100,000%
	15	100,000%		

Таблица 8. Динамика латинских слогов у Плавта по Ципфу

Динамика латинских слогов у Плавта по Ципфу, Первая цифра	Динамика. чисел	Значение *F(n)* или вероятность встретить цифру первой	Кумулятивные числа	Значение *F(n)* или вероятность встретить цифру первой
1	18	30,103%	18	30,10%
2	8	17,609%	26	47,712%
3	8	12,494%	34	60,206%
4	9	9,691%	43	69,897%
5	3	7,918%	46	77,815%
6	2	6,695%	48	84,510%
7	3	5,799%	51	90,309%
8	4	5,115%	55	95,424%
9	2	4,576%	57	100,000%
		100,000%		

Числа в исследованных произведениях изменяются от 9 (Потоки научно-технической информации) до 97 (Динамика патентов и рефератов в Chemical Abstracts (1907-2003).

Значение *F(n),* или вероятность встретить цифру первой во всех примерах составляют при числах 1-9 от 30,103% до 4,576%, а кумулятивные числа - от 30,103% до 100,000%.

Таким образом, закон Бенфорда применим к информатике (Information Sciences), т.е. ***подтверждается выдвинутая нами гипотеза и описана простыми алгебраическими уравнениями.***

Моделирование простыми алгебраическими уравнениями из информатики.

Перейдем к моделированию простыми алгебраическими уравнениями перечисленных примеров по информатике (рис. 4-12).

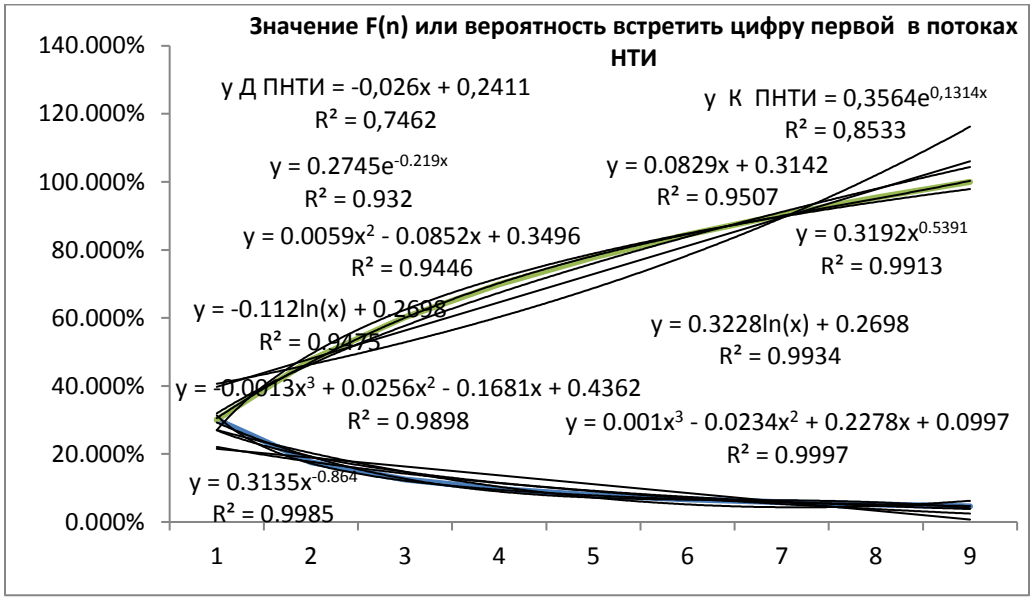

Рис.4. Значение *F(n)* или вероятность встретить цифру первой в потоках НТИ

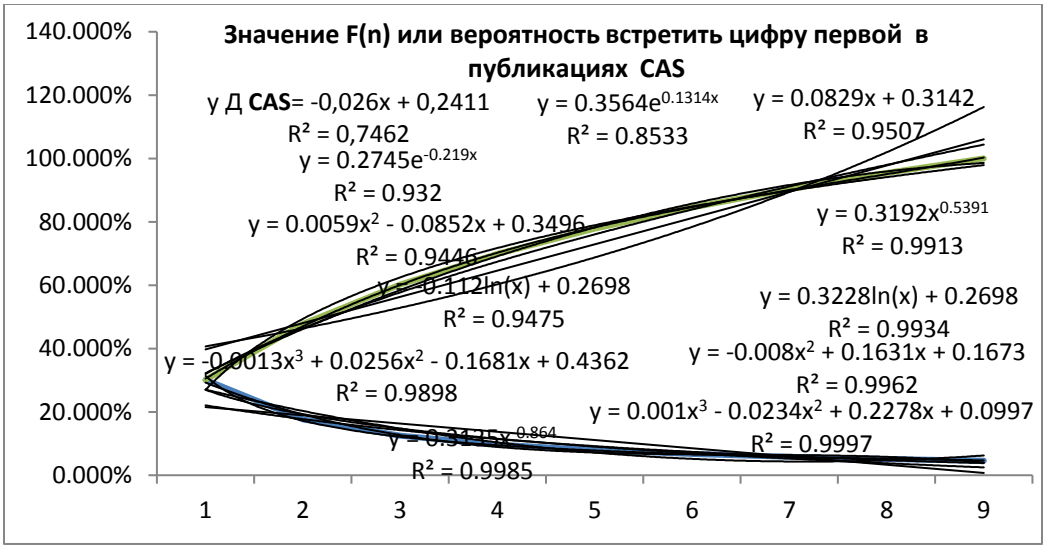

Рис.5. Значение *F(n)* или вероятность встретить цифру первой в публикациях CA

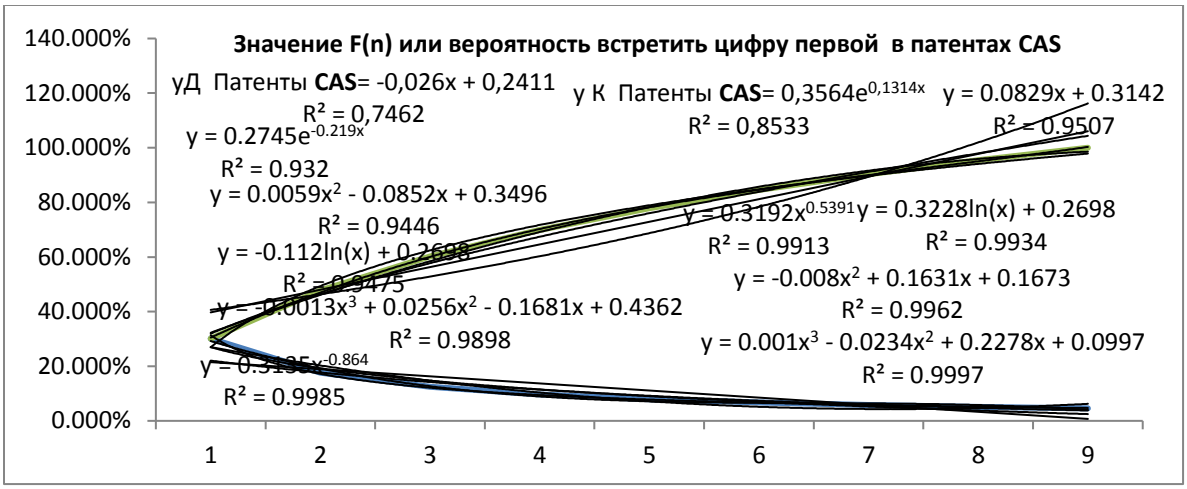

Рис.6 Значение *F(n)* или вероятность встретить цифру первой в патентах CAS

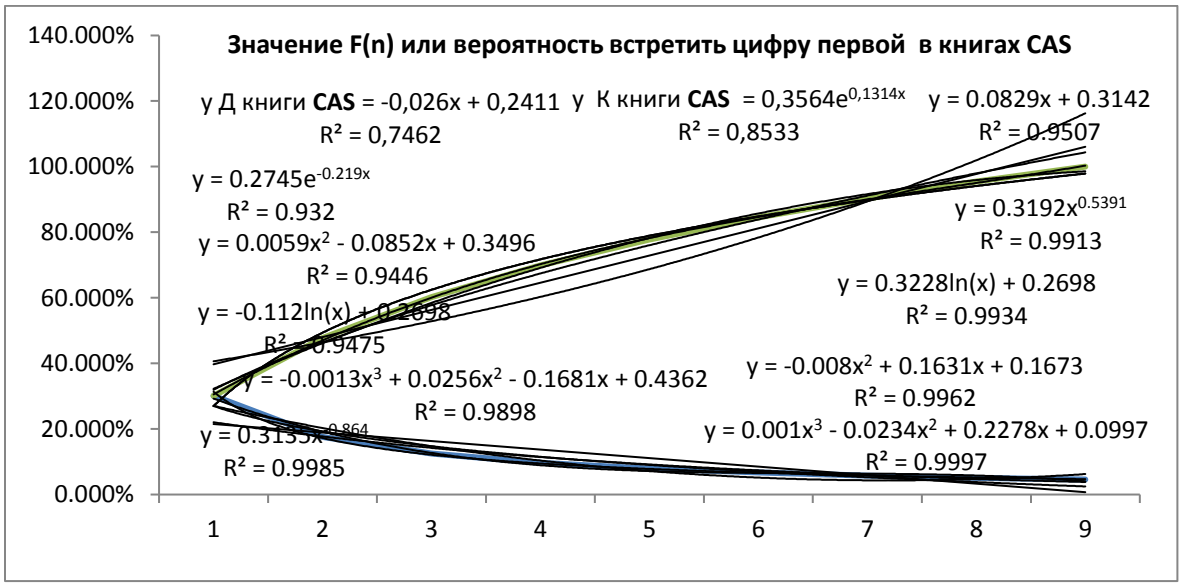

Рис.7. Значение *F(n)* или вероятность встретить цифру первой в книгах CAS

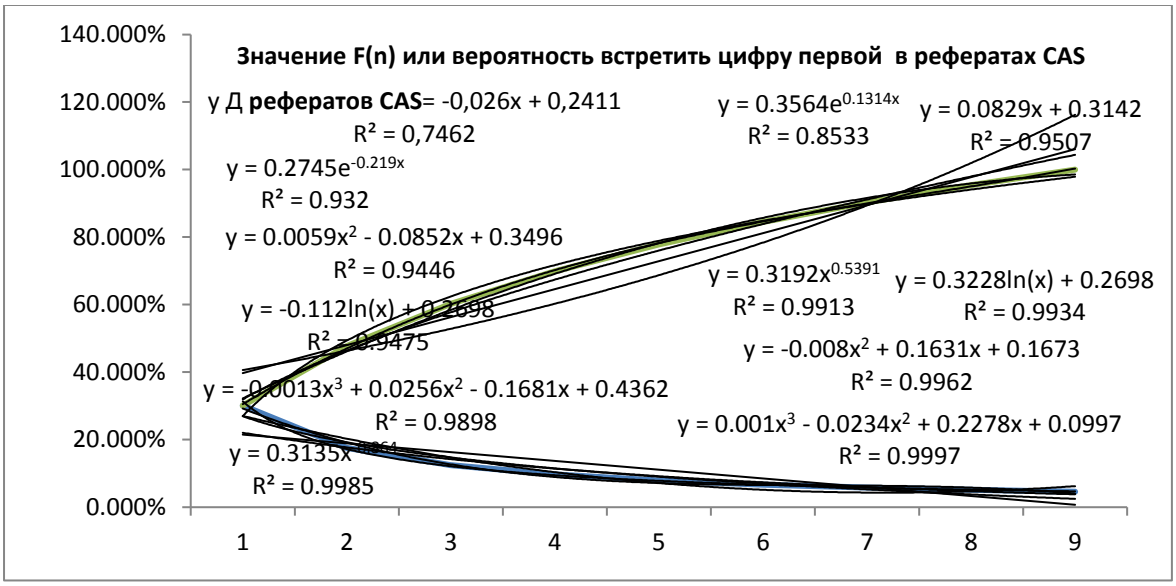

Рис.8. Значение *F(n)* или вероятность встретить цифру первой в рефератах CAS

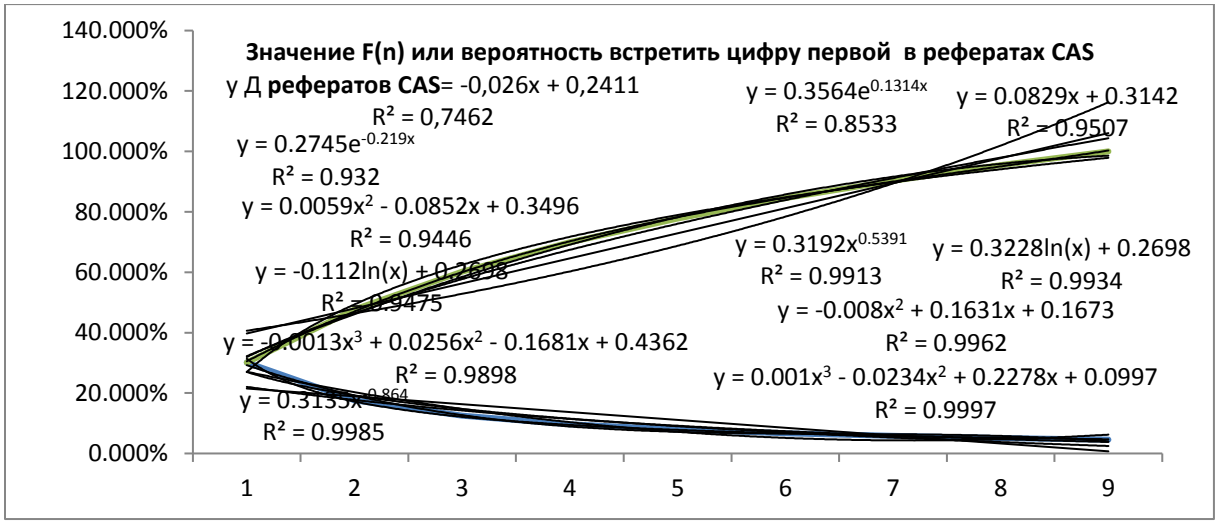

Рис. 9. Значение *F(n)* или вероятность встретить цифру первой в рефератах CAS

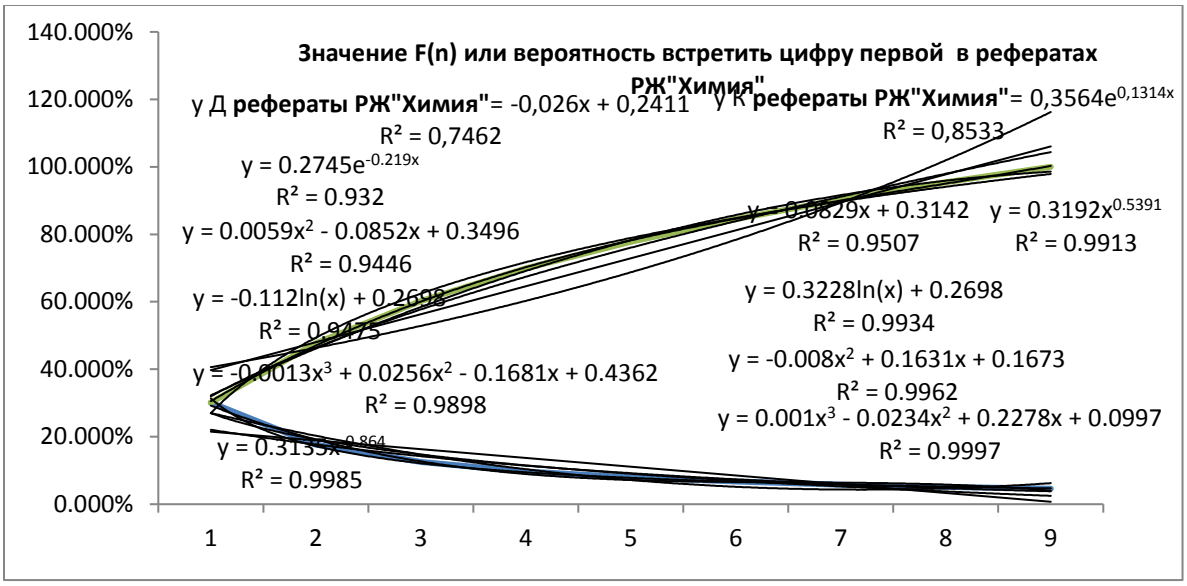

Рис.10. Значение *F(n)* или вероятность встретить цифру первой в рефератах РЖ «Химия" ВИНИТИ ПАН

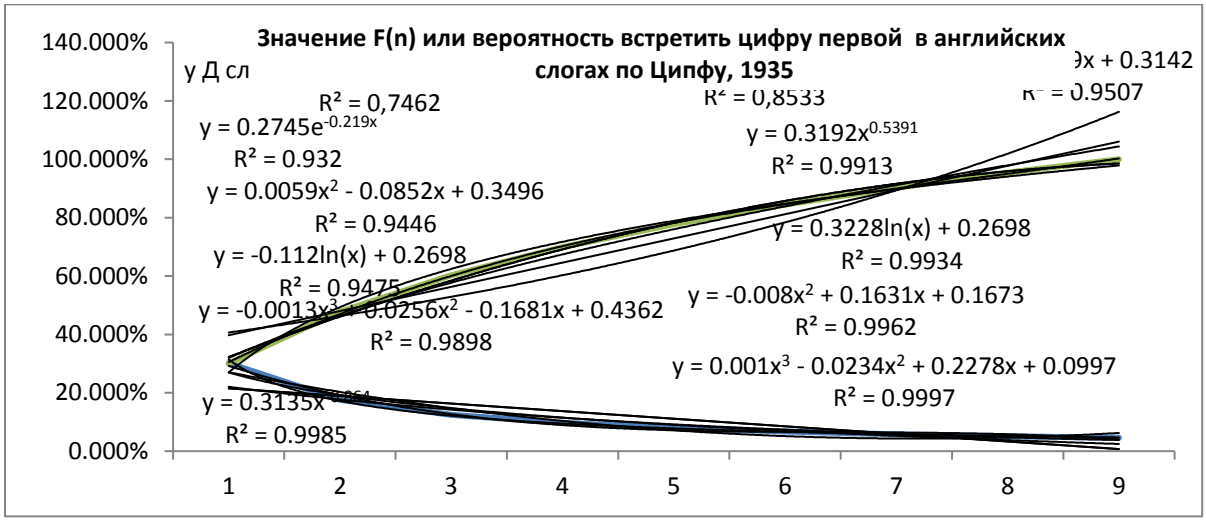

Рис.11. Значение F(n) или вероятность встретить цифру первой в английских слогах по Ципфу [6]

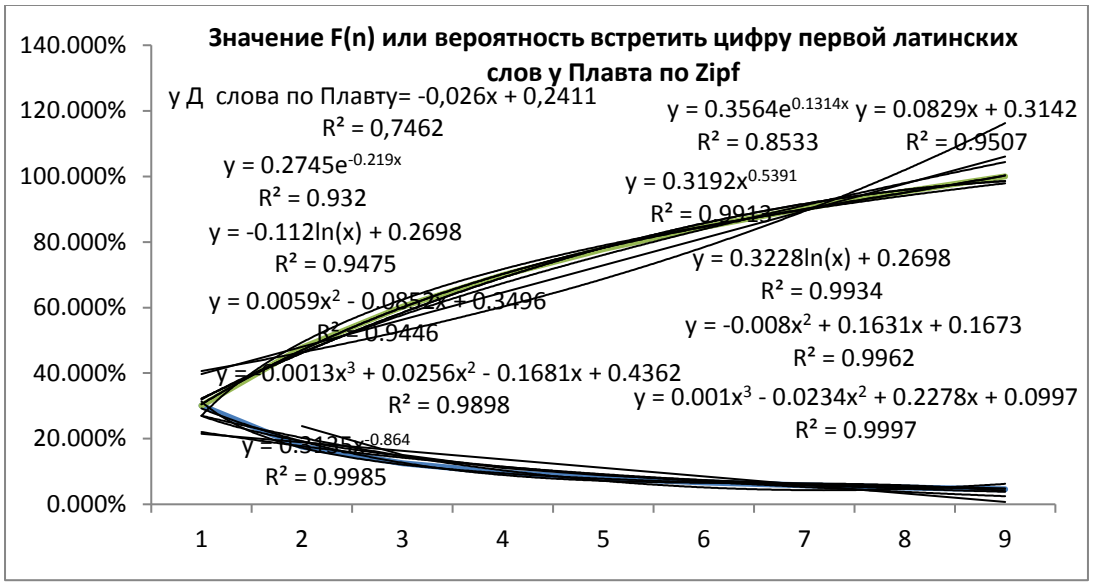

Рис.12. Динамика латинских слогов у Плавта по Ципфу [6]

Представленные простые алгебраические уравнения для всех примеров на рис. 4-12 динамики значений *F(n)* имеют одинаковые значения, например, $y_{\text{д ПНТИ}}$ = -0,026x + 0,2411, R² = 0,7462; $y = 0{,}2745e^{-0{,}219x}$, R² = 0,932, $y = 0{,}0059x^2 - 0{,}0852x + 0{,}3496$, R² = 0,9446, y = -0,112ln(x) + 0,2698, R² = 0,9475, $y = -0{,}0013x^3 + 0{,}0256x^2 - 0{,}1681x + 0{,}4362$, R² = 0,9898, $y = 0{,}3135x^{-0{,}864}$, R² = 0,9985 описывается с достаточной тояностью экспоненциальным уравнением полиномом второй степени, логарифмическимм уравнением, полиномом третьнй степени и степенным уравнением, а значения *F(n* для кумуляты ПНТИ; $y_{\text{к ПНТИ}} = 0{,}3564e^{0{,}1314x}$, R² = 0,8533; y = 0,0829x + 0,3142,R² = 0,9507; $y = 0{,}3192x^{0{,}5391,}$, R² = 0,991; y = 0,3228ln(x) + 0,2698, R² = 0,9934; $y = 0{,}001x^3 - 0{,}0234x^2$ +0,2278x + 0,0997, R² = 0,9997 описывается с достаточной тоснмтью линейным, степенным т лошгарифмическим уравнениями и полиномом третьей степени.,

Следует отметить, что *F(n)* или вероятность встретить цифру первой по простым алгебраическим уравнениям имеют различные значения: для динамики 0,2745=27,45% и для кумуляты 0,3564=35,64% (экспоненциальное уравнение), для динамики 0,3135=31,35% и для кумуляты 0,3192=31,92 (степенное уравнение) и для динамики 0,2411=24,11% и для кумуляты 0,3142=31,42 (линейное уравнение)

Таким образом, моделирование *F(n)* или вероятность встретить цифру первой по простым алгебраическим уравнениям подтверждает значения 30,108% по днамике и 0,3135=31,35% и по кумуляте 0,3192=31,92% (степенное уравнение) и по кумуляте 0,3142=31,42% (линейное уравнение) и описываются с наибольшей точностью от экспоненциального уравнения до полинома третьей степени. Наилучшей моделью является полином третьей степени.

Выводы

1 Впервые закон Бенфорда подтверждается для информатики и уточняется простыми алгебраическими уравнениями для динамики и кумулятивного числа цифр.
2 Впервые также показано по простым алгебраическим уравнениям различие динамики и кумуляты для относительной и относительной экспоненциальной скоростей изменения *F(n)* или вероятности встретить цифру первой.
3 Закон Бенфорда по лексикологии сближается с исследованными нами квантитативными характеристиками в области информатики и другими областями знания о Вселенной, т.е. является всемирным ·

References

1. Baguzin. *Zakon Benforda ili zakon pervoj cifry* [Benford law or first digit law] http://baguzin.ru/wp/zakon-benforda-ili-zakon-pervoj-tsifry/
2. Mlodinov L. *(Ne) sovershennaja sluchajnost'. Kak sluchaj upravljaet nashej zhizn'ju* [(Not) perfect coincidence. How chance controls our lives]. URL: http://baguzin.ru/wp/leonard-mlodinov-nesovershennaya-sluch/
3. Schetnikov A. I.. Shchetnikova A. V. Teaching and Researching Seminar "Distribution of the First Significant Digits" *Math. Ed., 2002, Issue 2(21), Pages 108–123* (Mi mo525)
4. *Ritera - kāto | Roshia shoseki senmon mise Nauka Japan* [Litera - Cart | Russian book specialty shop Nauka · Japan] URL: naukajapan.jp/detail.php?id=153315&PHPID.
5. *Klimov Ju.N. Kvantitativnaja leksikologija, korpusnaja lingvistika i kolichestvennaja informatika: monografija.* [Quantitative lexicology, corpus linguistics and quantitative informatics: monograph], Moscow, OchU VO "MMA", 340 p., hard. 2016. ISBN 9785904360542 R153315
6. Zipf G. K. Human behavior and the principle of least effort. «Addison-Wesley Publishing» Cambridge, 1949.